AN INVESTIGATION INTO THE MOLECULAR EPIDEMIOLOGY AND EVOLUTION OF HIV-1 AMONG NEWLY INFECTED PATIENTS IN HIGH AND LOW PREVALENCE SETTINGS

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy

by

John Christopher Ambrose

February 2014

ABSTRACT

Public health strategies to tackle HIV-1 epidemics in a variety of national settings need to be aware of the factors involved in transmission of infections. There is support in the literature for the hypothesis that individuals with recent infection may contribute disproportionately to onward transmission due to the high viral load and lack of infection status associated with this stage. This thesis sets out to further explore this risk group, and to develop methods to monitor its impact on epidemics.

Two epidemic settings, the United Kingdom and Kumasi, Ghana, were investigated using molecular epidemiological techniques to assess the role of individuals with recent infection in the formation of transmission clusters. A classifier of HIV-1 infection length was developed based upon the proportion of mixed nucleotides within consensus *pol* gene sequences and applied to phylogenies constructed using viral sequences obtained from each cohort. In the Ghanaian setting, the performance of the nucleotide ambiguity classifier was compared to an antibody avidity based measure of infection length, to gauge the usefulness of both approaches in a sub-Saharan setting. In order to more fully explore the complexity of intra-host HIV-1 quasi-species dynamics in the early phase of infection, a well defined cohort of UK-based individuals, some with multiple pre-treatment time points, had their virus deep-sequenced and analysed using a subpopulation reconstruction approach.

The proportion of recent infections identified within the UK HIV-1 epidemic by the classifier agreed well with previous studies. Application of the classifier to UK-wide phylogenies revealed disproportionate linkage of recently infected individuals to clusters across subtypes. Phylogenetic analysis of the Kumasi cohort did not reveal a highly clustered epidemic. Transmitted drug resistance was present at a level consistent with reports elsewhere in West Africa. Comparison of the antibody avidity based measure of infection length with the nucleotide ambiguity classifier indicated both markers co-segregate together, but produced differing estimates of the proportion of recent infections in the cohort. Deep-sequencing of recently infected individuals revealed some individuals may have been infected with more than one viral subpopulation, whilst others appeared to have been infected with a single subpopulation.

This work supports the utility of a consensus sequence based measure of infection length in assessing the role of recently infected individuals in driving epidemics on a large scale. Such a measure will need further refinement and validation depending on the setting used, but presents a potentially useful biomarker that could be used in conjunction with other clinical parameters. Deep-sequencing of HIV-1 in recently infected individuals points to the quasi-species complexity that exists between and within individuals, and the subpopulation reconstruction approach taken in this work reveals dynamics at play which have the potential to impact on vaccine design and molecular epidemiological monitoring of epidemics.

DECLARATION

This thesis is the result of my own work. The material presented here has not been presented and is not being presented, either wholly or in part, for any other degree or qualification. Some of the technical procedures were carried out in collaboration with individuals at the institutions below or elsewhere.

The research work was carried out at the Department of Virology, at the University College London Medical School, Royal Free Hospital, London, and at the Department of Clinical Infection, Microbiology and Immunology, Institute of Infection and Global Health, University of Liverpool

John Christopher Ambrose

ACKNOWLEDGEMENTS

Firstly, thanks must go to my supervisor, Anna Maria, for giving me the chance to undertake a PhD with her in the first place, for enabling me to pursue my geeky side in bioinformatics and my adventurous side in travels to Africa, and for her support and guidance over the last four years.

Next, I have to thank all the people I have worked with during my studies; at the Royal Free Hospital in London, the Institute of Infection and Global Health in Liverpool, and Komfo Anokye Hospital in Kumasi. I fear a list of names will only end up accidentally excluding some individuals – so hopefully people will know who they are, and will appreciate my gratitude for their being there and offering me much needed advice, assistance and/or friendship.

To my friends outside of my various workplaces, some of whom have undertaken PhD studies themselves, thanks has to be given to those who provided solicitude and comfort when things seemed like they weren't going so well, and who helped put things in perspective with timely and judicious application of alcohol and fun.

Finally, thanks to mum and dad, and my sister Jane (and the three newer additions to the family too) for being there whenever needed.

ABBREVIATIONS

°C	Degrees Centigrade
AIDS	Acquired Immune Deficiency Syndrome
APOBEC3G	Apolipoprotein B mRNA-Editing, Enzyme-Catalytic, Polypeptide-like 3G
ART	Antiretroviral Therapy
ARV	Antiretroviral
AUC	Area Under the Curve
AZT	Zidovudine
BED	HIV-1 Capture EIA Assay (uses branched peptide that includes gp41 immunodominant sequences from HIV-1 subtypes B, E, and D)
BLAST	Basic Local Alignment Search Tool
bp	Base Pair
CA	Capsid
CCR5	C-C chemokine receptor type 5
CD	Cluster of Differentiation
cDNA	Complementary deoxyribonucleic acid
CI	Confidence Interval
cm	Centimetre
COMET	COntext-based Modeling for Expeditious Typing
CRF	Circulating Recombinant Form
CTL	Cytotoxic T-lymphocyte
CXCR4	C-X-C chemokine receptor type 4
DC	Dendritic Cells
DNA	Deoxyribonucleic Acid

dNTP	Deoxyribonucleoside Triphosphate
DRAM	Drug Resistance Associated Mutation
ds	Double-stranded
EIA	Enzyme-Immuno Assay
ELISA	Enzyme-Linked Immunosorbent Assay
Env	Envelope
ER	Endoplasmic Reticulum
FPR	False Positive Rate
g	Gravitational Constant
GALT	Gut-Associated Lymphoid Tissue
GI	Gastrointestinal
GTR	General Time Reversible
h	Hour
HAART	Highly Active Anti-Retroviral Therapy
HIV	Human Immunodeficiency Virus
HLA	Human Leukocyte Antigen
IDU	Injecting Drug Users
IFN-	Interferon
Ig-	Immunoglobulin-
IL-	Interleukin-
IN	Integrase
IQR	Interquartile Range
IUPAC	International Union of Pure and Applied Chemistry
KATH	Komfo Anokye Teaching Hospital
kb	Kilobase

KIR	Killer-cell Immunoglobulin-like Receptor
KNUST	Kwame Nkrumah University of Science and Technology
L	Litre
LTR	Long Terminal Repeat
MA	Matrix
MCRA	Most Common Recent Ancestor
mg	Milligram
MID	Multiplex Identifier
mins	Minutes
ml	Millilitre
mM	Millimolar
mRNA	Messenger Ribonucleic Acid
MSM	Men who have Sex with Men
MUSCLE	MUltiple Sequence Comparison by Log-Expectation
NA	Not applicable/available
NC	Nucleocapsid
ng	Nanogram
NHS	National Health Service
NIH	National Institutes of Health
NK	Natural Killer
nM	Nanomolar
nm	Nanometre
NNRTI	Non-Nucleoside Reverse Transcriptase Inhibitor
NPV	Negative Predictive Value
NRTI	Nucleoside Reverse Transcriptase Inhibitor

OR	Odds Ratio
ORF	Open Reading Frame
PAMP	Pathogen Associated Molecular Pattern
PBMC	Peripheral Blood Mononuclear Cell
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain Reaction
PHI	Primary HIV Infection
PI	Protease Inhibitor
PIC	Pre-integration Complex
Pol	Polymerase
PPV	Positive Predictive Value
PR	Protease
PrEP	Pre-exposure Prophylaxis
PWID	People Who Inject Drugs
PWIJ	People Who Inject Drugs
RFH	Royal Free Hospital
RITA	Recent Infection Testing Algorithm
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristics
rpm	Revolutions per Minute
RT	Reverse Transcriptase
RT-PCR	Reverse Transcription Polymerase Chain Reaction
secs	Seconds
SGA	Single-genome Amplification
SIV	Simian Immunodeficiency Virus

SMH	St Mary's Hospital
SMRT	Single Molecule Real Time
SOD	Standard Optical Density
ss	Single-stranded
STD	Sexually Transmitted Disease
STI	Sexually Transmitted Infection
SU	Surface
TDR	Transmitted Drug Resistance
TM	Transmembrane
UK	United Kingdom
UKCHIC	UK Collaborative HIV Cohort
UKHIVDRD	UK HIV Drug Resistance Database
URF	Unique Recombinant Form
WHO	World Health Organisation
yr	Year
µg	Microgram
µl	Microlitre
µM	Micromolar

TABLE OF CONTENTS

LIST OF CONFERENCE PROCEEDINGS AND PAPERS IN PREPARATION

Conference proceedings:

Ambrose, J.C., Foster, G.M., Chaytor, S.C., Booth, C.L., Garcia-Diaz, A., Geretti, A. (2010). Population Sequence Nucleotide Ambiguity as a Measure of HIV-1 Infection Length (poster). 17th Conference on Retroviruses and Opportunistic Infections, San Francisco, USA.

Ambrose, J.C. (2011). The Role of Recent Infection in Driving the UK HIV-1 Epidemic (oral). 15th Annual Resistance and Antiviral Therapy Meeting, London, UK.

Ambrose, J.C. (2011). Using Population Sequence Nucleotide Ambiguity to Estimate HIV-1 Infection Length (poster). 18th HIV Dynamics and Evolution, Galway, Ireland.

Ambrose, J.C. (2011). The Contribution of Recent Infections to HIV-1 Transmission Clusters in the UK (oral). 18th HIV Dynamics and Evolution, Galway, Ireland.

Ambrose, J.C. (2012). Is recent infection a major source of onward HIV transmission in the UK? (oral).16th Annual Resistance and Antiviral Therapy Meeting, London, UK.

Publications in preparation based upon work carried out during the term of this candidature:

Ambrose, JC; Garcia-Diaz, AM; Mackie, NE; Fidler, S; Porter, K; Geretti, AM; UK HIV Drug Resistance Database: An investigation into the extent to which individuals with recent infection drive the formation of HIV-1 transmission clusters in the UK

(sequence data was obtained from the UK HIV DRD, together with sequence and avidity data from the Royal Free Hospital generated by A Garcia-Diaz and other members of the diagnostic staff, sequence and clinical data from St Mary's Hospital supplied by N Mackie and S Fidler, and sequence and clinical data from K Porter at the UK Register of HIV Seroconverters; J Ambrose and AM Geretti designed the study; J Ambrose carried out the analysis)

Ambrose, JC; Kenny, JG; Hall, N; Geretti, AM: HIV-1 subpopulation dynamics in recent infection and over time

(J Kenny and N Hall provided advice and represent the Centre of Genomic Research which carried out the 454 pyrosequencing; J Ambrose and AM Geretti designed the study; J Ambrose carried out the generation of the PCR products prior to sequencing and carried out the analysis of the data)

Ambrose, JC; Appiah, L; Phillips, R; Geretti, AM: Molecular epidemiology of new HIV-1 diagnoses in Kumasi, Ghana

(L Appiah assisted in collection of blood samples and clinical information; R Phillips and AM Geretti designed the study; J Ambrose performed sequencing and avidity assays and carried out the data analysis)

Chapter One

General introduction

1 General introduction

1.1 HIV

In mid-1981 a number of homosexual men in New York and Los Angeles presented with unusual opportunistic infections such as Pneumocystis carinii pneumonia and aggressive Kaposi's sarcoma (Friedman-Kien, Laubenstein et al. 1981; Gottlieb, Schroff et al. 1981). All individuals were found to have a deficit in the $CD4^+$ T cell subset of their immune system. Similar presentations in previously healthy individuals were reported in other risk groups and in other countries around the world, particularly in the Caribbean, Western Europe and Africa (Masur, Michelis et al. 1982; Rozenbaum, Coulaud et al. 1982; Vilaseca, Arnau et al. 1982; Clumeck, Mascart-Lemone et al. 1983; Harris, Small et al. 1983; Pitchenik, Fischl et al. 1983). By late 1982 the term Acquired Immune Deficiency Syndrome (AIDS) was being used to describe the condition, and by 1983, evidence that the condition was linked to a retrovirus was published (Barré-Sinoussi, Chermann et al. 1983). The virus we now refer to as Human Immunodeficiency Virus was finally linked to AIDS in 1984 (Levy, Hoffman et al. 1984).

The Human Immunodeficiency Virus can be divided into two major types, HIV-1 and HIV-2. HIV-2, although spread in the same way as HIV-1, seems to be less virulent and transmissible, and hence is not as widespread as HIV-1 in the global epidemic (De Cock, Adjorlolo et al. 1993). HIV-1 and HIV-2 are lentiviruses within the retrovirus family that infects human immune system cells such as T-helper cells, macrophages and dendritic cells, and if left untreated lead to Acquired Immune Deficiency Syndrome (AIDS) through destruction of the adaptive immune system, making the body vulnerable to opportunistic infections such as tuberculosis, and cancers such as Kaposi's sarcoma and AIDS-related non-Hodgkin's lymphoma (Embretson, Zupancic et al. 1993; Fauci 1993). Perhaps circulating in

humans as early as 1910 (Sharp and Hahn 2011), an estimated 35.3 million people were living with HIV (type 1 and 2) world-wide in 2012, with AIDS related deaths having reached an estimated 36 million deaths over that period (UNAIDS 2013; UNAIDS 2013).

1.2 Biology of HIV-1

1.2.1 Structure of the virion

Mature HIV-1 virions are variable in size, and have been shown to have diameters ranging between 125 and 145nm (Briggs, Wilk et al. 2003; Briggs, Grünewald et al. 2006). They are comprised of an envelope of host cell derived lipid bi-layer containing an average of 14 (range 4-35) trimer molecules of the transmembrane gp41 and gp120 glycoproteins used for host cell binding and fusion (Klein and Bjorkman 2010). Within the envelope is a capsid (p24) core organised into a fullerene-type cone structure made up of a sheet of ~1500 capsid molecules organised into a hexagonal lattice comprised of approximately 250 hexameric rings organised into a cylindrical conformation, capped at the wide end by 7 pentameric rings, and at the narrow end by 5 rings (Arhel 2010; Jiang, Ablan et al. 2011). The core contains further structural proteins derived from the Gag and Gag-Pol polyproteins, namely: matrix (p17), associated with the inner virion membrane; nucleocapsid (p7), which coats the RNA molecules and has a role in RNA dimerization and packaging (Kafaie, Song et al. 2008); and p6, which enables Vpr incorporation into the mature virion (Kondo, Mammano et al. 1995); together with the accessory proteins Vif, Vpr and Nef, and the protease, reverse transcriptase and integrase enzymes proteins and two positive strands of the HIV-1 RNA genome (Fig 1.1.).

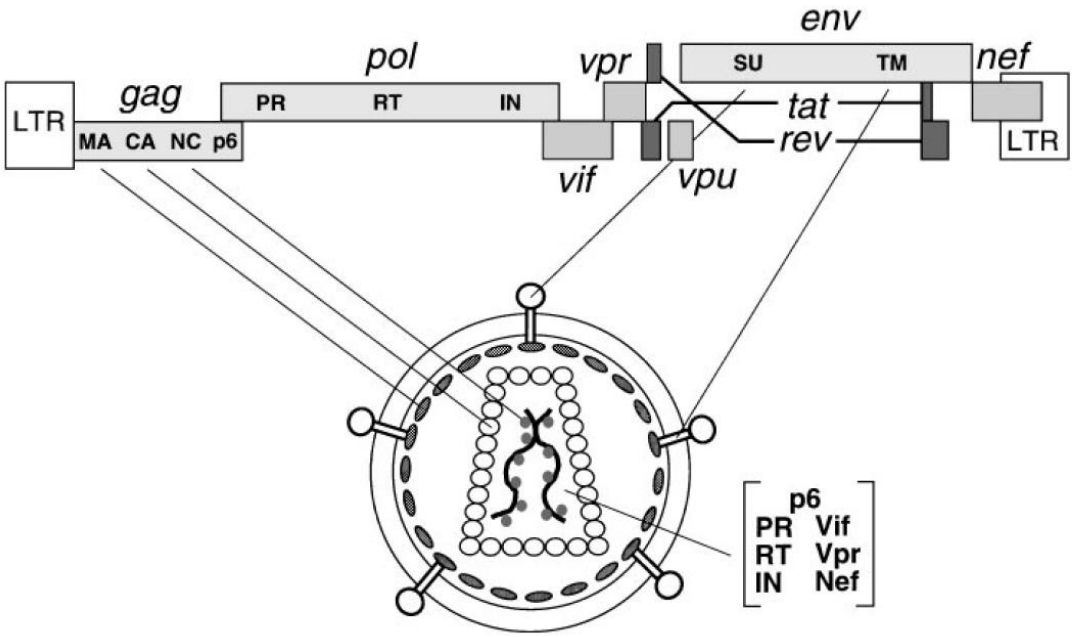


Figure 1.1. Taken from (Frankel and Young 1998). The genomic and virion structure of HIV-1.

1.2.2 Genome

The HIV-1 genome is an RNA positive strand genome of approximately 9.7kb in length, containing 15 genes. Six of these genes encode structural proteins (Table 1.1), four of which are derived from the Gag polyprotein p55 to form the core structural proteins: matrix (p17), capsid (p24), nucleocapsid (p7) and p6; with the remaining two being derived from the Env gp160 polyprotein to form the envelope structural proteins: surface (gp120) and transmembrane (gp41) (Frankel and Young 1998) (Fig 1.2.). Three enzymes are derived from the Gag-Pol precursor polyprotein; protease, reverse transcriptase and integrase. The remaining six genes encode accessory proteins: three of which are found in the HIV-1 virion (Vif, Vpr, Nef), two of which are the regulatory proteins Tat and Rev, and one of which is Vpu, which has an indirect role in virion assembly. The HIV-1 genome is flanked at both ends by Long Terminal Repeat (LTR) regions approximately 640bp in length, which are essential for integration of the viral genome into the host DNA, and contain elements that

interact with a variety of potential host cell transcription factors (Berkhout, Gatignol et al. 1990; Pereira, Bentley et al. 2000).

HIV Genes and Gene Products		
GAG		
MAtrix	**p17**	Targets Gag and Gag-Pol precursor polyproteins to the plasma membrane prior to viral Aids incorporation of env glycoproteins with long cytoplasmic tails into viral particles Facilitates infection of nondividing cell types, principally macrophages
CApsid	**p24**	Forms the core of the virus particle
NucleoCapsid	**p7**	Coats the genomic RNA inside the virion core (protects it from nucleases and compacts it within the core)
p6	**p6**	Important for incorporation of Vpr during viral assembly
POL		
PRotease	**PR**	Cleaves polyprotein at several sites to produce MA, CA, NC, and p6 proteins from Gag and PR, RT, and IN proteins from Pol
Reverse Transcriptase	**RT**	Catalyzes both RNA-dependent and DNA-dependent DNA polymerization reactions and contains an RNase H domain that cleaves the RNA portion of RNA-DNA hybrids generated during the reaction
INtegrase	**IN**	Catalyzes a series of reactions to integrate the viral genome into a host chromosome
ENV		
SUrface	**gp120**	Binds CD4 with high affinity and plays key role in attachment to specific cell surface receptors (three gp120s and gp41s combine in a trimer of heterodimers to form the envelope spike)
TransMembrane	**gp41**	Mediates fusion between the viral and cellular membranes following receptor binding
Regulatory		
Trans-Activator of Transcription	**TAT**	Binds to TAR RNA element and activates transcription initiation and elongation from the LTR promoter
Regulator of Expression of Virion proteins	**REV**	Binds to Rev Responsive Element and promotes nuclear export, stabilization, and utilization of the viral mRNAs containing RRE
Accessory		
Viral Infectivity Factor	**VIF**	Promotes infectivity but not production of viral particles by disrupting the antiviral activity of the human virus restriction factor APOBEC
Viral Protein R	**VPR**	Incorporated into the virion and may aid in targeting nuclear import of preintegration complexes, cell growth arrest, transactivation of cellular genes, and induction of cellular
Viral Protein U	**VPU**	Plays a role in i) degradation of CD4 in the endoplasmic reticulum, and ii) enhancement of virion release from the plasma membrane of HIV-1-infected cells
NEgative Factor	**NEF**	Downregulates CD4, the primary viral receptor, and MHC class I molecules, increases viral infectivity
HIV Genomic Structural Elements		
Long Terminal Repeat	**LTR**	DNA sequence flanking the genome of integrated proviruses that contains important regulatory regions, such as transcription initiation and polyadenylation sites
TransActivation Response element	**TAR**	Target sequence for viral transactivation, the binding site for Tat protein and cellular proteins
Rev Responsive Element	**RRE**	RNA element encoded within the env region of HIV-1 which contains a high affinity site for REV

Table 1.1. List of HIV-1 genes, gene products, and functions.

1.2.3 Replication cycle

HIV-1 targets three main host immune cell types: $CD4^+$ T-lymphocytes, macrophages and dendritic cells (Boggiano and Littman 2007; Lackner and Veazey 2007). This somewhat limited tropism may be due to the cell surface receptors that are targeted by the virion – the CD4 receptor and a co-receptor that is usually the chemokine receptor CCR5 in early infection, although the CXCR4 co-receptor is often targeted later in HIV-1 infection as the virus evolves. Each trimer on the surface of the virion is comprised of three heterodimers of the surface gp120 glycoprotein non-covalently linked to the transmembrane gp41 glycoprotein, which is in turn anchored to the viral envelope membrane.

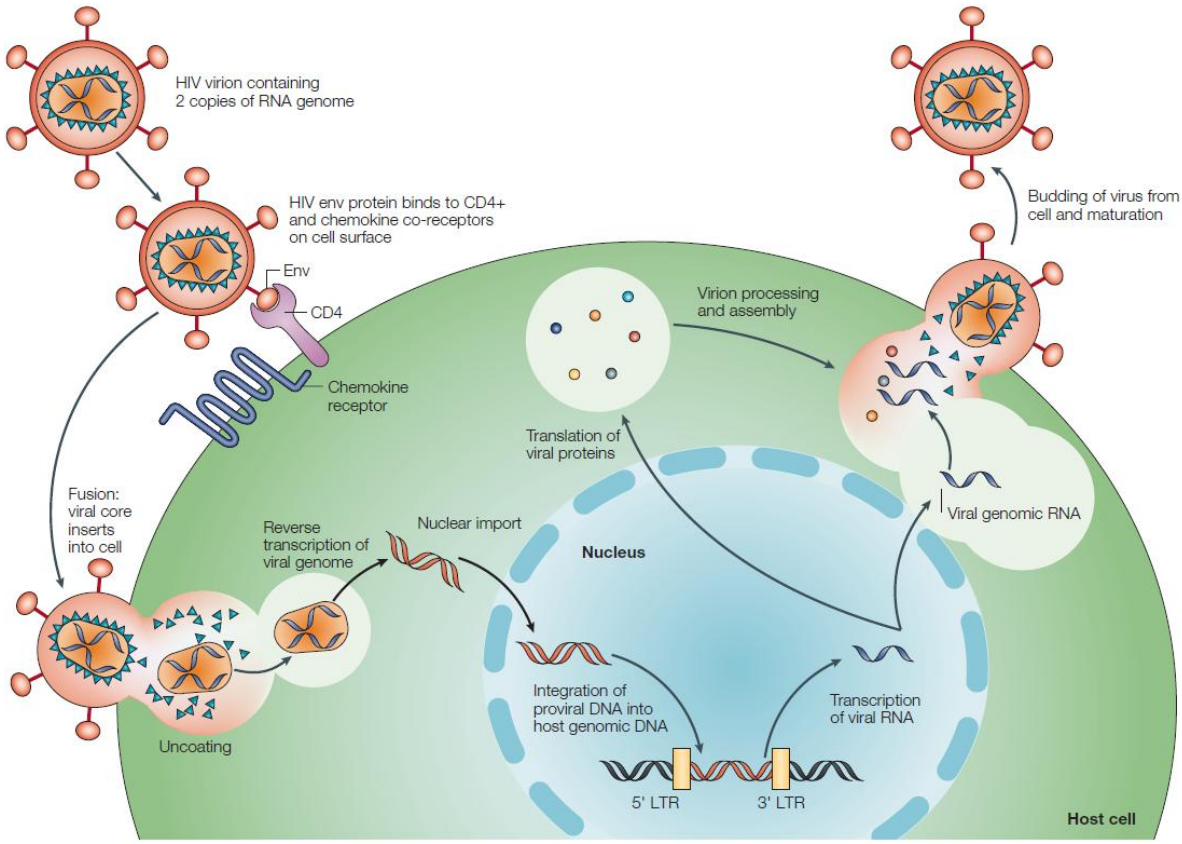


Fig 1.2. Taken from (Rambaut, Posada et al. 2004). Life-cycle of HIV-1.

Binding: The CD4 receptor binds between the inner and outer domains of the gp120 molecule (Briz, Poveda et al. 2006) and leads to a conformational change involving the V1/V2 and V3 variable loops, which leads to the exposure of the co-receptor binding site (Clapham and McKnight 2001; Engelman and Cherepanov 2012) (Fig 1.2.).

Fusion: Once the co-receptor has been bound, the N-terminal domain of gp41 is exposed, and the hydrophobic fusion peptide is inserted into the host cell membrane (Briz, Poveda et al. 2006) and a series of conformational rearrangements between trimerized N-terminal and C-terminal heptad repeat motifs generates a six helix bundle, which causes a hairpin structure in the gp41 molecules and the virion and host cell membranes to be brought into contact and fuse (Melikyan, Markosyan et al. 2000; Engelman and Cherepanov 2012).

Uncoating: Once virion-cell fusion has occurred, the viral core is released into the cell interior and uncoating occurs (Fig 1.2.). The precise timing of the uncoating is still to be fully elucidated, with several competing models being proposed (Arhel 2010; Fassati 2012). There is evidence to suggest that uncoating can begin as little as 30-45 minutes post-fusion, when the core is released into the cytoplasm. However, there is also evidence to suggest that if uncoating occurs too early, reverse transcription is inhibited (Forshey, von Schwedler et al. 2002). It also appears that initiation of reverse transcription itself has an influence on the timing of the uncoating process (Hulme, Perez et al. 2011). Other evidence suggests that uncoating does not occur close to the cell surface but takes place in a gradual, stepwise manner during transport of the core to the nucleus, in a process possibly instigated by conformational changes in the core complex as a result of the initiation of reverse transcription together with interactions with the host cell machinery. Another model proposes that uncoating does not initiate until the core has reached the nuclear membrane and reverse transcription is complete, whereupon the capsid lattice is disassembled and the pre-

integration complex (PIC) is transferred across a nuclear pore into the interior of the nucleus (Arhel 2010).

Integration: The pre-integration complex is made up of the double stranded reverse transcribed viral genomic DNA and integrase. Other proteins have also been shown to aggregate with the PIC, such as matrix and reverse transcriptase, but there is good evidence that integrase alone is necessary and sufficient to integrate viral DNA into the host genome (Farnet and Haseltine 1991; Bukrinsky, Sharova et al. 1993). Inside the nucleus, the PIC must become associated with host genomic DNA in order to integrate, and the efficiency of this process appears to be increased by association with the host cell transcriptional co-activator Lens-Epithelial-Derived Growth Factor (LEDGF/p75) (Christ and Debyser 2013), which helps to target the viral DNA to active transcription units within the host genome, an evolutionary strategy that may enable a more rapid production of viral RNA before the host cell is targeted for destruction by the immune system, which can occur as quickly as a few days after the cell is infected (Craigie and Bushman 2012). The integration reaction proceeds in a series of steps starting with removal of two nucleotides from the 3′ ends of the viral DNA, which then react with phosphodiester bonds on opposing strands of the host chromosomal DNA and bond with the 5′-phosphates (Craigie and Bushman 2012; Engelman and Cherepanov 2012). The 3′ ends of the viral DNA are subsequently joined to the host DNA covalently and the remaining nucleotide gaps and overhangs are repaired by host DNA repair machinery (Craigie and Bushman 2012).

Transcription: The 5′ LTR region of the HIV-1 integrated provirus contains a number of potential transcription binding sites that can be targeted by a number of cellular transcription factors such as NF-κB, Sp1, and TBP (Jones and Peterlin 1994). Transcription is carried out by the host RNA polymerase II, but requires the trans-activator of transcription (Tat) accessory protein to drive elongation of the transcript. Tat binds to a highly conserved RNA

hairpin structure, termed the transactivation response (TAR) element, and recruits cellular host factors such as the human Positive Transcription Elongation Factor b (P-TEFb), which acts to recruit further host factors to the RNA polymerase II in order to modulate transcription elongation (Ott, Geyer et al. 2011). The Regulator of Expression of Virion proteins (Rev) protein is required to enable transport of singly spliced and unspliced HIV-1 mRNA transcripts out of the nucleus. Initially, the first viral mRNAs produced by transcription are predominantly doubly spliced and code for the Tat, Rev, and Nef proteins. The production of the Rev protein allows singly and unspliced mRNA to be exported from the nucleus and translated into the additional polyproteins required for virion formation, or packaged into the immature virion as the complete HIV-1 genome. Export mediated by Rev requires the mRNA to contain a Rev Response Element (RRE), which is present as a 351bp sequence within the Env coding region of partially and unspliced viral mRNA (Malim, Hauber et al. 1989; Mann, Mikaélian et al. 1994).

Assembly and budding: Assembly takes place at the host cell plasma membrane once all virion components have been translated and transported to the membrane along with the two capped and polyadenylated RNA genomes to be packaged inside the virion. The Gag structural proteins are present as the unprocessed Gag polyprotein, but this is sufficient for virion assembly to begin and for budding to occur. Assembly of an immature virion particle takes ~10 minutes on average, and budding is enabled by interaction of Gag with the Endosomal Sorting Complexes Required for Transport (ESCRT) machinery normally used during mitosis and vesicle formation (Usami, Popov et al. 2009). Once budding has occurred, the virion particle can mature through proteolysis of the Gag and Gag-Pol polyproteins by protease to form the full complement of proteins and enzymes required by the virion. The production of these proteins leads to major conformational changes in the virion, to produce a mature, fully infectious particle (Sundquist and Kräusslich 2012).

1.2.4 Latency and reactivation

Even if viral replication is suppressed to below detectable levels through antiretroviral therapy, HIV-1 persists in the resting $CD4^+$ T-cell reservoir of infected individuals as proviral DNA integrated into the cellular genome (Siliciano, Kajdas et al. 2003; Laird, Eisele et al. 2013). This reservoir can remain in the body for the remainder of an infected individual's lifetime because of the natural function of this immune cell subset, and can be reactivated into fully replication competent virus when antiretroviral drug pressure is removed (Wong, Hezareh et al. 1997; Chun and Fauci 2012).

1.2.5 Cell-cell and cell-free intra-host transmission

HIV-1 can make contact and infect host cells via two mechanisms: cell-free transmission, and cell-to-cell transmission. Transmission of virions across virological synapses formed between infected and uninfected $CD4^+$ T-cells have been observed (Jolly and Sattentau 2004; Piguet and Sattentau 2004), and this mode of transmission could represent a more efficient transmission mechanism than cell-free transmission, as it leads to a concentrated targeting of multiple virions at one site on a donor cell, which occurs in close proximity to cell membrane receptors on the target cell, as opposed to diffusion of cell-free virions, with no certainty of encountering a suitable target cell-type (Sundquist and Kräusslich 2012). Indeed, it may be the case that host antiretroviral restriction factors reduce the ability of HIV-1 virions to be released from the host cell, instead anchoring them to the membrane in a manner which further promotes cell-to-cell transfer (Jolly, Booth et al. 2010).

1.3 Natural history of infection

1.3.1 Founding virus theory, quasi-species

To infect a new host, HIV-1 virions must gain access and entry to the relevant target cells of the host immune system. The major routes for this to occur are penile-vaginal and penile-

rectal sex, mother to child transmission, and injecting drug use. Of these, the primary route worldwide is vaginal-penile sex (Hladik and McElrath 2008; UNAIDS 2013). As such, it is the genital epithelia that present the first line of defence HIV-1 must circumvent in the majority of transmissions so as to establish infection. Transmission can occur both from male to female and female to male during vaginal-penile sex, and hence the epithelia of both the penis and vagina/cervix have been investigated in terms of target cell availability and infection rates (Haase 2010; Anderson, Politch et al. 2011). It is predominantly investigations of vaginal transmission using non-human primate models, together with work on individuals with recent infection, that has provided convincing evidence for a genetic bottleneck at the mucosal barrier, or soon after passing through the mucosal barrier, that results in the infection in the new host being founded by in some cases as few as one or two virions from the donor inoculum (Zhang, MacKenzie et al. 1993; Zhu, Mo et al. 1993; Keele, Li et al. 2009) (Fig 1.3.). This may be a direct result of the mucosal barrier acting to block virion passage, or a consequence of other factors once the virions have overcome this initial hurdle (Boeras, Hraber et al. 2011; Parrish, Gao et al. 2013), although recent work casts doubt on the cervical mucosa being the barrier that generates the transmission bottleneck (Shen, Ding et al. 2012).

Evidence from follow up of individuals diagnosed with early infection supports the hypothesis that this initially highly homogenous virus goes on to increase in diversity in a relatively linear manner over the first years of infection (prior to initiation of therapy) through host immune system pressure and the high replication and error rate of replication in HIV-1 infection (Zhang, MacKenzie et al. 1993; Zhu, Mo et al. 1993; Shankarappa, Margolick et al. 1999; Keele, Giorgi et al. 2008). There is limited data on the extent to which this increase in diversity occurs equally across the genome, and the impact co-infection and subsequent recombination may have on genetic diversity. The latter consideration may be important in the context of infections founded by multiple virions, particularly in penile-rectal sex and

injecting drug use associated infections, where there is good evidence the majority of infections are founded by more than one virion (Bar, Li et al. 2010; Li, Bar et al. 2010). In chapter five of this thesis, this topic is investigated using deep-sequencing (see section 1.8) in a cohort of patients diagnosed with early infection.

The increase in diversity of the founding virus in early infection generates a cloud, or 'quasi-species', of related but distinct viruses, These viruses can be thought of as occupying – and continuously exploring – a sequence space, where viruses with different mutations occupy different parts of this space, and have corresponding differences in fitness (Lauring and Andino 2010). Formal mathematical definitions of quasi-species do exist, and although it is debatable the extent to which these definitions are applicable to viral populations, the concept of a diverse but interacting swarm of viruses, under constant selective pressure by the host immune system and/or drug pressure, may still provide useful insights into intra-host viral evolution (Eigen 1971; Domingo 2002; Moya, Holmes et al. 2004; Lauring and Andino 2010).

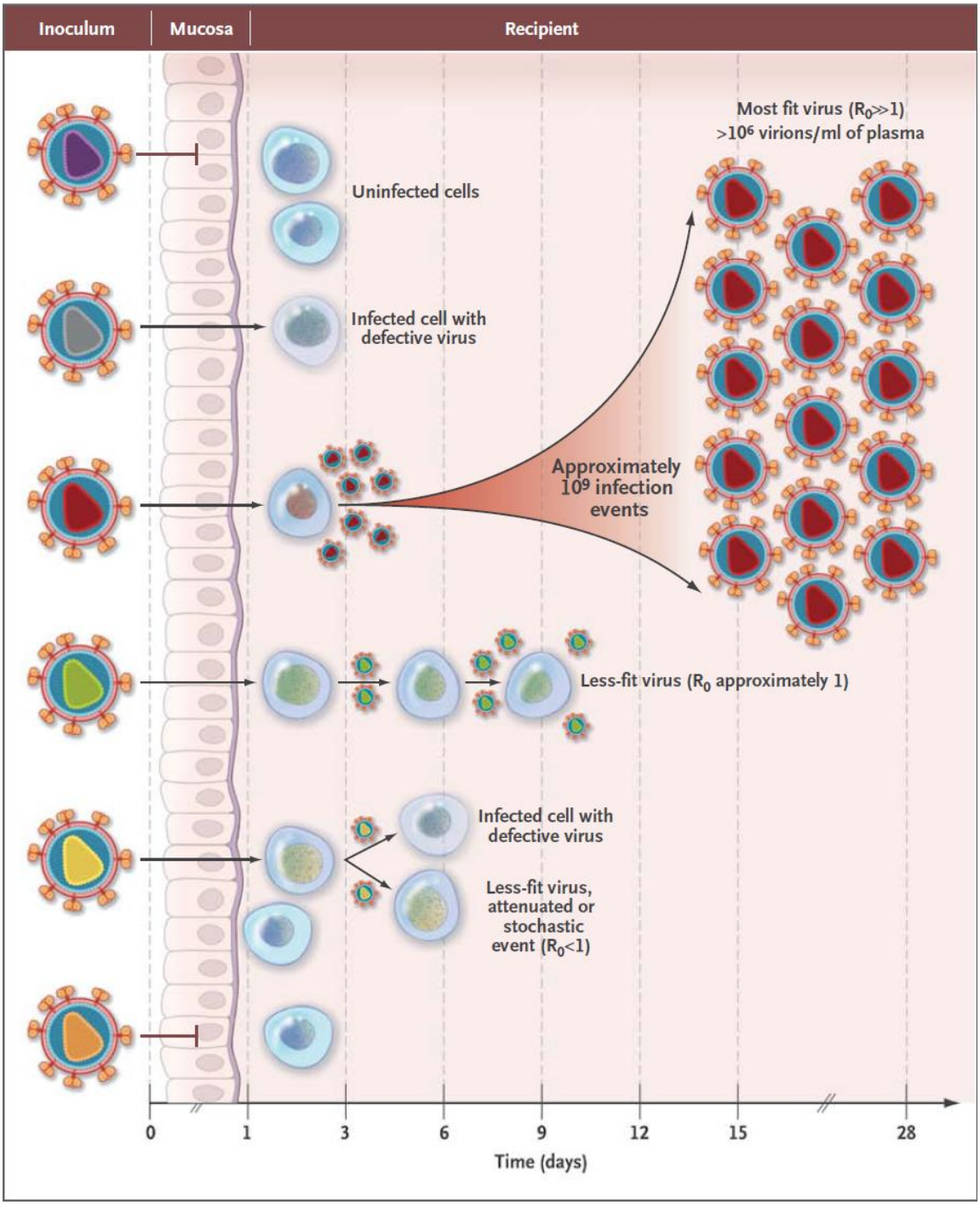


Figure 1.3. Taken from (Cohen, Shaw et al. 2011). Model for HIV-1 transmission which illustrates the theoretical process behind single/limited founder viruses in HIV-1 infection. (R_0 for an infection is the number of onward cases one infection generates on average during its infectious period in a totally susceptible population. Where $R_0 < 1$, the infection tends to die out over time, where $R_0 > 1$, the infection may become widespread in the population).

1.3.2 Viral load in blood, GALT and genital compartments

Regardless of the route of entry into the new host, there is a period of ~10 days, termed the 'eclipse' phase, when the virus replicates in the individual in the absence of symptoms, and no virus is detectable in the blood (Palmer, Wiegand et al. 2003; Sickinger, Jonas et al. 2008). HIV-1 RNA levels ("viral load") in plasma peak in the second to third week after infection and then start to decline with the emergence of HIV-specific immune responses. The key processes and sites of initial infection and replication during this eclipse phase are still not completely understood. Vaginal transmission offers the most sophisticated human tissue explant and non-human primate models available, and it is therefore the model that is the best understood (Miller, Alexander et al. 1989; Miller, Li et al. 2005; Hladik and Hope 2009; Keele, Li et al. 2009). $CD4^+$ and Langerhans cells are highly likely to be the primary targets of infection and replication upon immediate incursion by the virus, in both vaginal and penile submucosa (Cohen, Shaw et al. 2011). Replication and further infection of these target cells leads to rapid dissemination of the virus to lymph nodes and the gut associated lymphoid tissue (GALT), before spreading to other compartments and tissues. The GALT is the focus of the main $CD4^+$ T cell loss in the initial stages of infection and throughout infection prior to therapy, and this loss is predominantly from the $CCR5^+$ $CD4^+$ T cell subset (Brenchley, Schacker et al. 2004). In terms of within-host compartmentalisation of the virus and onward transmission, there is good evidence that in early stages of infection the viral load in the male genital compartment lags behind the viral load in the blood, and peaks around 4 weeks post-infection, but is controlled by around week 10, and then displays stable levels as the patient moves to chronic infection. In general, the viral load in the semen is lower than in the blood, and together with other evidence, such as discordant viral phenotype distributions and the lack of a strong correlation between viral RNA level in semen and $CD4^+$ cell count in blood, points to compartmentalisation of the virus (Coombs, Speck et al. 1998; Pilcher, Shugars et

al. 2001), a process which is likely to be occurring in the female genital tissues also (Craigo and Gupta 2006).

1.3.3 Recent and established infection

The dynamics of early infection lead to large changes in viral load within the infected individual prior to reaching the relatively stable levels seen in chronic infection. Evidence suggests that the time taken for virions to penetrate the mucosal barrier, reach relevant cells and gain entry can be short, but that it may be up to a week later before the infection begins to spread through the lymphatic system and reach the lymph nodes and GALT (Hladik and McElrath 2008; Cohen, Shaw et al. 2011). Beyond this time scale the virus begins to disseminate through the body and to lay down the reservoir of proviral DNA that will remain with the patient through infection, whether viral replication is suppressed through therapy or not (Chun, Engel et al. 1998). The initial phase of infection can usefully be divided up into stages defined by various biological markers (Fiebig, Wright et al. 2003) (Fig 1.4.), although these stages are open to adaptation to new, more sensitive assays as they become available (Ananworanich, Fletcher et al. 2013).

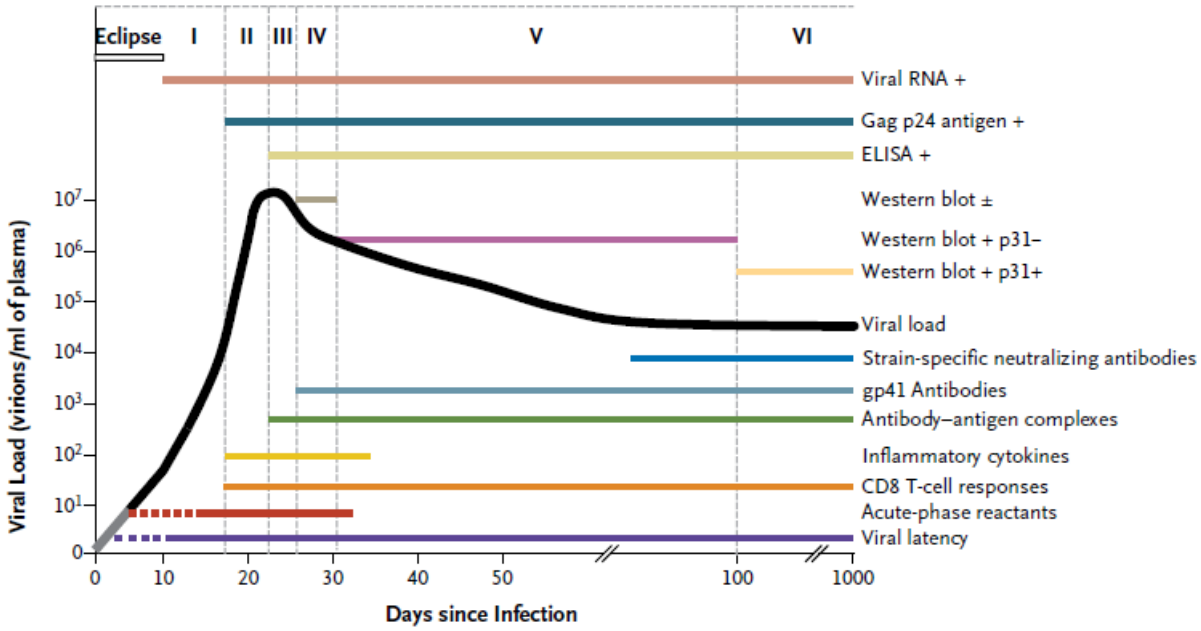


Figure 1.4. Taken from (Cohen, Shaw et al. 2011). Natural history of HIV-1 infection, showing the Fiebig infection staging system in terms of the results of standard clinical laboratory tests.

After the initial eclipse phase, the first detectable signs of HIV-1 infection, the time point termed Fiebig stage I, can be obtained by carrying out sensitive viral RNA detection methods on patient blood, generally from day 11 after infection (Fig 1.4.). Fiebig stage II is defined by the detectability of the p24 gag protein antigen, associated with a rise of viral load to 10000 copies/ml and above. It is around this point in infection that peak viremia is reached, which can be as high as 10^7copies/ml (Pilcher, Joaki et al. 2007), and can be linked to the severity of seroconversion symptoms experienced around this time (Lavreys, Baeten et al. 2002; Kelley, Barbour et al. 2007). Approximately 5 days later, Enzyme-Linked Immunosorbent Specific Assays (ELISA) can begin to detect immunoglobulin M antibodies against recombinant HIV antigens. Fiebig stages IV and V are defined by the results of a Western blot assay (protein immunoblots) that detects antibody reactivities against individual viral proteins (typically p24 (from *gag*), p31 (from *pol*) and gp41 or gp120/160 (from *env*). Stage IV is defined by an

equivocal/indeterminate Western blot result, with stage V being an unequivocally positive Western blot result but p31 integrase antibody negative. The final, open-ended stage, Fiebig stage VI, is defined as Western blot positive and p31 integrase antibody positive (Fig 1.4.).

The initial immune response to HIV-1 infection may be mediated by detection of viral specific markers by receptors evolved to detect Pathogen Associated Molecular Patterns (PAMPs) such as double or single stranded RNA and particular viral proteins (Mogensen, Melchjorsen et al. 2010). Activation of molecular pattern specific receptors, such as the Toll-like receptors (TLR) 7 and 8, can prompt the production of immunomodulatory cytokines such a interferon-α and interleukin-15 (Chang and Altfeld 2010). There is also a detectable increase in molecular markers of cell apoptosis (Gasper-Smith, Crossman et al. 2008). The cytokine and chemokine surge can play a role in speeding up the dissemination of HIV-1 infection by i) summoning immune cells to the site of infection, thereby increasing the number of target cells available for virion entry, and by ii) increasing inflammation in the infected locale, leading to a breakdown of tight junctions between epithelial cells (Chang and Altfeld 2010; Nazli, Chan et al. 2010). The chemokine response is likely to play a role in increasing the natural killer (NK) cell response in the region of infection. This cell type is highly heterogeneous in terms of its ability to target specific pathogens, due to differences in its killer-cell immunoglobulin-like receptors (KIRs), and there is evidence that individuals with more favourable KIRs seem to have a better ability to control HIV-1 infection (Martin, Gao et al. 2002; Martin, Qi et al. 2007). Initial antibodies within the adaptive immune response primarily target gp41, the trunk of the envelope trimer, but in a non-neutralising manner. $CD8^+$ T cell response peaks around 1-2 weeks after peak viremia, and immediately leads to exploration of CTL escape mutants at various key epitope sites along the HIV-1 genome, primarily in *env* and *nef* (McMichael, Borrow et al. 2009).

1.3.4 Detecting recent infection

Of major importance in terms of onward transmission of HIV-1 within a population is the high viral load associated with acute HIV-1 infection prior to being brought under partial immune control, both in blood and genital compartments (Pilcher, Joaki et al. 2007; Morrison, Demers et al. 2010). This elevated viral titre, together with other factors such as lack of awareness of infection status, may mean that acute/recent infections account for a disproportionate number of onward transmissions compared to the established phase of infection (Jacquez, Koopman et al. 1994; Koopman, Jacquez et al. 1997; Hollingsworth, Anderson et al. 2008). To gain a better understanding of the incidence of HIV-1 and the role of recent infection in epidemics, a number of assays have been developed to detect recent infection in newly diagnosed individuals. These are primarily based upon measuring increasing levels of antibodies, such as the branched peptide antigen (BED) based capture Enzyme-Immuno Assay (EIA), or measurements of antigen-antibody binding strength that rely upon the principle of antibody hypermutation, and the corresponding maturation of antibody affinity to HIV-1 antigens (Parekh, Kennedy et al. 2002; Chawla, Murphy et al. 2007; Mastro, Kim et al. 2010). Such assays are increasingly combined together with clinical biomarkers such as $CD4^+$ cell count, into algorithms such as the Recent Infection Testing Algorithm used by Public Health England (Garrett, Lattimore et al. 2012) and the Serologic Testing Algorithm for Recent HIV Seroconversion (STARHS), developed by the US Centers for Disease Control and Prevention (CDC) (Janssen, Satten et al. 1998). It is important to note that the majority, if not all of these assays, have been developed in Western settings, and have been optimised predominantly on subtype B infections. As such, it is unclear to what extent they are immediately appropriate for use in other settings, such as sub-Saharan Africa, without further validation (Sakarovitch, Rouet et al. 2007).

A number of groups have also investigated the utility of HIV-1 sequence information in estimating infection length, be it using Sanger *pol* sequences obtained from patients as part of routine antiretroviral drug resistance surveillance (Kouyos, von Wyl et al. 2011; Ragonnet-Cronin, Aris-Brosou et al. 2012; Andersson, Shao et al. 2013), or deep-sequencing data obtained through second-generation sequencing platforms such as 454 pyrosequencing (Giorgi, Funkhouser et al. 2010; Poon, McGovern et al. 2011). A further investigation of the use of HIV-1 sequence information to determine infection length forms chapter three of this thesis.

1.4 Epidemiology of HIV-1

1.4.1 Subtypes

HIV-1 can be divided into 4 major groups, each representing a separate zoonotic transmission event from chimpanzees or gorillas to humans, probably in the course of bushmeat trade and consumption (Buonaguro, Tornesello et al. 2007; Sharp and Hahn 2011; Vallari, Holzmayer et al. 2011). Group M, for 'main', is the major group, which is responsible for the majority of HIV infections worldwide (Ariën, Abraha et al. 2005). Group M itself contains a diverse collection of HIV-1 strains, which cluster together into related clades. These 9 clades, labelled A–D, F–H, J and K, maintain their distinct branching patterns on phylogenetic trees regardless of which section of the genome is used for analysis, and have been termed subtypes (Fig 1.5.).

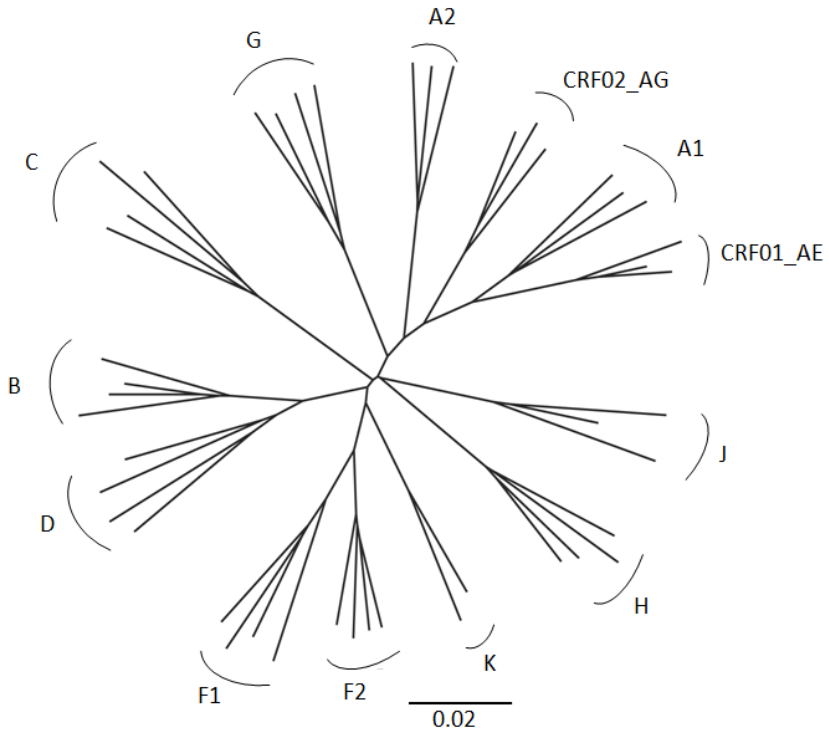


Figure 1.5. Group M subtype *pol* phylogenetic tree generated using Los Alamos Group M sequence reference alignment (http://www.hiv.lanl.gov/). Sub-subtypes are also shown. Scale bar is in nucleotide substitutions per site.

Further circulating strains of HIV-1 group M have been identified that have been found to map to different subtype clades depending on which part of the genome is used to draw phylogenies (Robertson, Sharp et al. 1995) – these appear to be the result of recombination between different strains of virus, made possible by co-infection with two or more different subtypes or recombinants of HIV-1, together with the process of template swapping between the two RNA genomes packaged in the HIV-1 virion during reverse transcription (Onafuwa-Nuga and Telesnitsky 2009). The HIV-1 nomenclature has been modified to take these strains into account, and they are termed circulating recombinant forms (CRFs) if they fulfil the criteria of being identified in three or more individuals not directly epidemiologically connected. To date, over 50 CRFs have been identified (Li, Ning et al. 2013; Foster, Ambrose

et al. 2014), and many more unique recombinant forms (URFs, recombinants that have only been identified in one or two epidemiologically unconnected individuals).

1.4.2 Origins of HIV-1

In terms of the diversity of HIV-1 and the distribution of URFs, Cameroon, the Republic of Congo, and the Democratic Republic of Congo have been found to have the highest number of complex forms of the virus. This fits with phylogenetic analysis of the relatedness of HIV-1 and Simian Immunodeficiency Viruses isolated from chimpanzees and gorillas from a variety of locations throughout central and West Africa, that suggests the origin of the different HIV-1 groups is likely to be Cameroon, the Republic of Congo, and/or the Democratic Republic of Congo (Van Heuverswyn and Peeters 2007; Sharp and Hahn 2011). Further evidence suggests that the zoonotic transmission event responsible for the group M pandemic took place sometime around the early 1900s (Korber, Muldoon et al. 2000). It is possible that zoonotic transmissions of SIV to humans had taken place many times prior to this through exposure to chimpanzee blood products during slaughter and consumption of bushmeat, but that the virus was not able to replicate and transmit itself to a sufficient extent to allow adaptation to its new host before being removed by the host immune system (Peeters, Courgnaud et al. 2002; Sharp and Hahn 2011). It has been speculated that the development of cities such as Léopoldville/Kinshasa, and the concomitant large influx of works and sex workers, may have been the factor that enabled a simian immunodeficiency virus to be serially passed through several human hosts in quick succession, giving it the opportunity to adapt to the new host immune system (Locatelli and Peeters 2012). Other theories propose that iatrogenic practises occurring through sleep-sickness and syphilis vaccination programs may have been the crucial factor in enabling transmission of the nascent pathogen between individuals (Marx, Alcabes et al. 2001). In comparison to the complexity of viral forms around the geographic cradle of HIV-1 group M virus, the global

HIV-1 group M subtype distribution is relatively simple in its major trends, and is largely geographical, probably due to a combination of founder effects and differences in transmissibility/fitness between subtypes (Sharp and Hahn 2011) (Fig 1.6.).

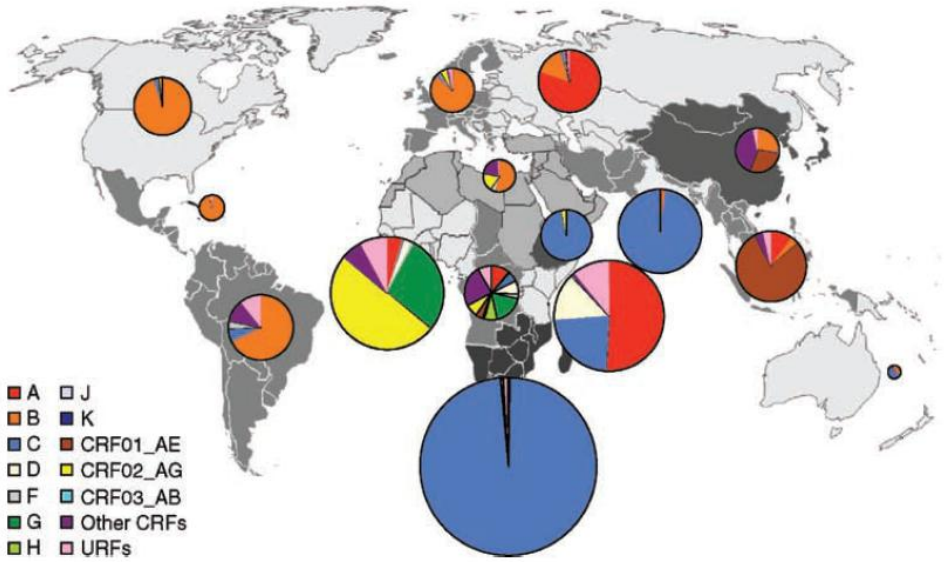


Figure 1.6. Taken from (Hemelaar, Gouws et al. 2011). Global HIV-1 subtype/CRF distribution 2004-2007. The surface areas of the pie charts correspond to the relative numbers of people living with HIV-1 in each particular region.

Of importance, HIV-1 subtype B predominates in Europe and North America, where the resources are available to spend on research and development, but subtype B is not the most globally prevalent in terms of the number of individuals infected worldwide (Fig 1.6.), potentially leading to a mismatch in terms of the effectiveness of diagnostic assays, antiretroviral treatments, and vaccine designs (Apetrei, Loussert-Ajaka et al. 1996; Peeters, Toure-Kane et al. 2003; Geretti 2006; Geretti, Harrison et al. 2009). Some of these issues are encountered in chapter six of this thesis, which deals with HIV in a West African setting.

1.4.3 Risk groups within the UK epidemic

In the early 1980s in the United Kingdom, the virus was initially spread within the men who have sex with men (MSM) community, and within networks of people who inject drugs (PWID), both initially subtype B virus epidemics (Robertson, Bucknall et al. 1986; Anonymous 1988; Brown, Lobidel et al. 1997). However, even as early as the mid 1990s, additional subtypes were circulating, and the epidemic was clearly involving heterosexual sex, and import of infections from outside of the United Kingdom (Clewley, Arnold et al. 1996; Brown, Lobidel et al. 1997). The proportion of non-B subtypes, and the contribution to the epidemic from different risk groups, continued to develop in the United Kingdom, and by the early 2000s the majority of new diagnoses or AIDS deaths were in the heterosexual risk group, and of these, over three-quarters were in individuals who were probably infected in Africa with non-B subtypes (The UK Collaborative Group for HIV and STI Surveillance 2007). Several lines of evidence demonstrate that the UK epidemic has experienced crossover between the different risk group 'sub-epidemics', with potential implications for clinical management and evolution of novel forms of the virus through recombination (Aggarwal, Smith et al. 2006; Fox, Castro et al. 2010), which could mirror the global increase in the contribution of recombinant forms to the pandemic (Hemelaar, Gouws et al. 2011).

1.5 Antiretroviral management of HIV-1 infections

1.5.1 History of drug development

The first clinical trial of an antiretroviral drug to combat HIV-1 infection, Zidovudine, was carried out in 1987 (Fischl, Richman et al. 1987), followed by investigations into Didanosine, Stavudine and other antiretroviral compounds that all acted by inhibiting action of the reverse transcriptase by mimicking nucleotides and terminating synthesis of the cDNA strand of the

HIV-1 genome (Balzarini, Herdewijn et al. 1989; Yarchoan, Mitsuya et al. 1989). The effectiveness of administering these drugs individually or in dual combinations was also investigated (Hammer, Katzenstein et al. 1996). However, it was the addition of a third drug, with a different mechanism of action, that was found to have major effects on the disease progression of individuals infected with HIV-1 (Collier, Coombs et al. 1996; D'Aquila, Hughes et al. 1996; Staszewski, Miller et al. 1996; Gulick, Mellors et al. 1997; Hammer, Squires et al. 1997). These results introduced the era of highly active antiretroviral therapy (HAART), which has continued to evolve as new drugs have been developed, but the principle of which is to target HIV-1 with drugs that have at least two distinct mechanisms of action, so as to more effectively contain the ability of HIV-1 to evolve mutations that would allow it to escape suppression by one or two drugs (Perelson 2002; Arts and Hazuda 2012).

1.5.2 Drug classes and mechanisms of action

There are currently 25 drugs for the treatment of HIV-1 licensed in the European Union (NAM 2013). 7 of these are combination pills, containing three or more drugs, whilst the remainder fall into 6 categories of mechanisms, each targeting different aspects of the HIV-1 replication cycle: Nucleoside and Nucleotide Reverse Transcriptase Inhibitors (NRTIs), Non-nucleoside Reverse Transcriptase Inhibitors (NNRTIs), Protease Inhibitors (PIs), Fusion Inhibitors, CCR5 Co-receptor Antagonists, and Integrase Strand Transfer Inhibitors (INSTIs).

Nucleoside or Nucleotide Reverse Transcriptase Inhibitors:

This class of antiretrovirals was the first to be developed, and acts by termination of viral DNA elongation (Mitsuya, Weinhold et al. 1985; Yarchoan, Weinhold et al. 1986). Nucleoside or nucleotide analogues enter the host cell and undergo phosphorylation by host kinases. These nucleotide analogues then compete with the host nucleotides for incorporation into the DNA chain. When a nucleotide analogue is incorporated it terminates elongation of

the chain due to its lack of a 3′-hydroxyl group on the deoxyribose portion of the molecule (Fig 1.7). Chain termination can occur during both RNA- or DNA-dependent DNA synthesis, interfering with production of either the forward or reverse strand of proviral DNA (Clavel and Hance 2004; Arts and Hazuda 2012).

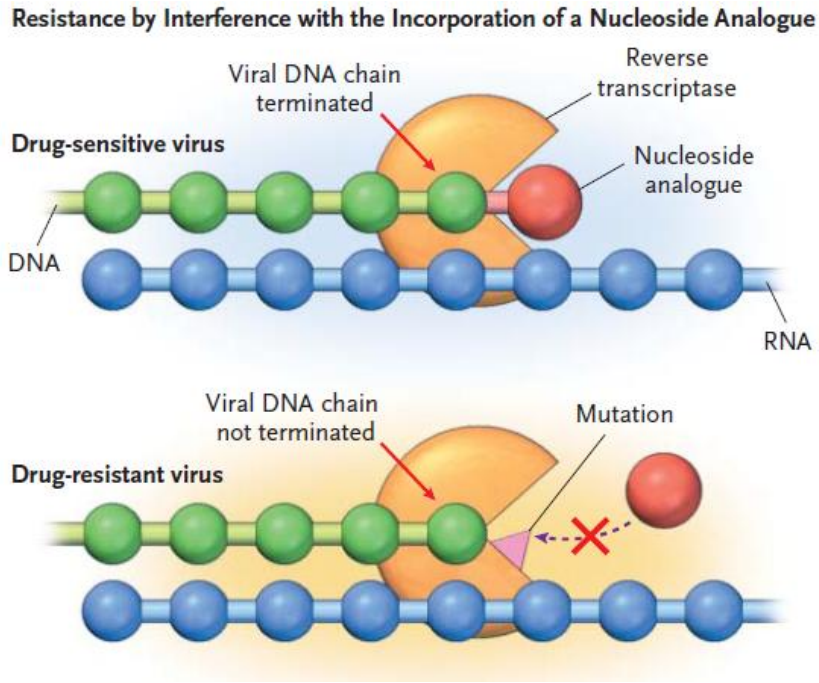


Figure 1.7. Taken from (Clavel and Hance 2004). Nucleoside or Nucleotide Reverse Transcriptase Inhibitor mechanism of action.

Non-nucleoside Reverse Transcriptase Inhibitors:

This class of molecule also targets the reverse transcription step of viral replication, but acts to inhibit the reverse transcriptase enzyme by non-competitive binding to a pocket close to the active site, thereby producing a conformational change in structure that reduces the activity of the enzyme (de Béthune 2010).

Protease Inhibitors:

This class of drug aims to block cleavage of the HIV-1 polyproteins (predominantly gag-pol), and thereby prevent production of mature virions, by competitive binding to the protease enzyme.

Entry Inhibitors:

These can be subdivided into Fusion Inhibitors and CCR5 Co-receptor Antagonists. The former, of which enfuvirtide is the only currently approved member of the class, are peptide-like molecules that block the conformational change of gp41 required to bring together and allow fusion of the viral envelope with the host cell membrane. The drug mimics specific sites within the gp41 structure (Kilby, Hopkins et al. 1998). The latter, of which maraviroc is the only currently approved class member, are small molecule inhibitors of CCR5 co-receptor binding that bind to the transmembrane region of the co-receptor causing allosteric inhibition of the interaction between the extracellular region of the co-receptor and the V3 region of the viral gp120 envelope protein. As HIV-1 can also use the CXCR4 co-receptor and occasionally other co-receptors for entry, it is necessary to determine the absence of CXCR4 tropic virus prior to administering CCR5 co-receptor antagonists (Soriano, Perno et al. 2009; Vandekerckhove, Wensing et al. 2011).

Integrase Strand Transfer Inhibitors:

These compounds act to target the strand transfer reaction involved in HIV-1 proviral DNA integration by binding to the complex formed between integrase and the pre-integrated viral DNA and interfering with the integrase active site binding to the DNA molecule (Pommier, Johnson et al. 2005).

1.5.3 Triple therapy

It is now standard to administer antiretroviral drugs to patients in combinations of usually three drugs, from at least two different classes. This was found early on to be more likely to result in long term suppression of the virus (Hammer, Katzenstein et al. 1996; Gulick, Mellors et al. 1997; Hammer, Squires et al. 1997). The basis for this lies in the high turnover of HIV-1 replication within a patient, together with the error prone nature of viral replication. Studies of viral populations, in combination with mathematical models of viral turnover and mutation rate, lead to the conclusion that the quasi-species of HIV-1 in an untreated individual is large enough that, statistically speaking, viral genomes with mutations conferring resistance to specific drugs already pre-exist in the patient prior to treatment (Coffin 1995), although many of these variants are likely to be non-viable. These mutants generally display reduced fitness in the absence of drug selective pressure, and are therefore often present at a low level within the quasi-species as usually single and less commonly double mutants (triple mutants are rare). Once drug pressure is introduced, if viral replication is not effectively suppressed, these pre-existing drug-resistant variants have the potential to become the dominant species as they outgrow other quasi-species members who lack such an evolutionary advantage (Coffin 1995; Nowak, Bonhoeffer et al. 1997). The presence of three drugs, requiring a much more unlikely combination of pre-existing mutations on the same viral genome, is the theoretical basis for combination therapy – particularly as some classes require multiple mutations to occur in the viral genome before they become ineffective (Clavel and Hance 2004).

1.5.4 Drug resistance

Even amongst antiretroviral drugs of the same class, there are differences in the type and number of mutations required by the virus to confer resistance – differences in the genetic barrier to resistance, with some drugs requiring only one mutation in the virus for them to

become ineffective whist others need a combination of mutations to occur in a specific order (Clavel and Hance 2004) (Table 1.2.). Viral load testing, where resources are available, allows monitoring of the effectiveness of viral suppression by the therapy being offered to a particular individual. When virological failure occurs, often through adherence issues (Yerly, Kaiser et al. 1999), it may lead to a switch in therapy, where one or two components of the therapy are replaced by other drugs with differing resistance profiles. Careful monitoring of potential drug resistance is required to ensure the correct therapy is initiated, and that any resistance that may have formed during virological failure to the previous regimen is unlikely to impact upon the effectiveness of the new regimen (Tang and Pillay 2004; Booth, Garcia-Diaz et al. 2007). Where drug resistance has developed in an individual, this can often be archived within viral reservoirs, which poses a long term risk of drug resistance if the patient is switched back to particular components of the previous regimen. It can also lead to the transmission of drug resistant virus to treatment naive individuals, which may limit treatment options available to them (Booth and Geretti 2007; Li, Paredes et al. 2013). Where such cases of transmitted drug resistance are picked up, suboptimal therapies can be avoided in preference of antiretrovirals to which the virus does not have any resistance, however, such genotypic screening for resistance is not available in all settings where antiretroviral therapy is available. More cross-sectional epidemiological monitoring of particular populations could give indications of the general level of circulating drug resistant virus, particularly in more resource limited populations where antiretrovirals are becoming more available, but where routine monitoring of drug resistance in patients failing therapy is not available (Ji, Li et al. 2011; Dudley, Chin et al. 2012; Gupta, Jordan et al. 2012).

Drugs	Mechanisms of Action	Mechanisms of Resistance
Nucleoside analogues Zidovudine Stavudine Lamivudine Didanosine Zalcitabine Abacavir	Analogues of normal nucleosides Active as triphosphate derivatives Incorporated into nascent viral DNA Prematurely terminate HIV DNA synthesis	Thymidine analogue mutations promote ATP-mediated and pyrophosphate-mediated excision of the incorporated terminator M184V or Q151M complex mutations impair incorporation of nucleoside analogues
Nucleotide analogues Tenofovir	Same as nucleoside analogues	K65R impairs incorporation of tenofovir into DNA Thymidine analogue mutations often associated with cross-resistance to tenofovir
Non-nucleoside reverse-transcriptase inhibitors Nevirapine Efavirenz Delavirdine	Bind a hydrophobic pocket of HIV type 1 reverse transcriptase Block polymerization of viral DNA Inactive against HIV type 2	Mutations reduce affinity of the inhibitors for the enzyme Single mutations generally sufficient to induce high level of resistance
Protease inhibitors Saquinavir Ritonavir Indinavir Nelfinavir Amprenavir Lopinavir	Structure derived from natural peptidic substrates of the HIV type 1 protease Bind the active site of the protease	Mutations reduce affinity of the inhibitors for the enzyme High-level resistance requires accumulation of mutations
Fusion inhibitors Enfuvirtide	36-Amino-acid peptide derived from the HR2 domain of glycoprotein 41 Interferes with glycoprotein 41–dependent membrane fusion	Mutations affect HR1, a domain of glycoprotein 41 whose interaction with HR2 promotes membrane fusion

Table 1.2. Adapted from (Clavel and Hance 2004). Antiretroviral agents used in the treatment of HIV-1 infection and their associated pathways of resistance.

1.6 Molecular epidemiology

1.6.1 Background theory: models of nucleotide evolution, maximum likelihood/Bayesian approaches

Information contained in viral sequences can provide useful insights into the timing of disease emergence and the composition of ancestral strains of the pathogen that can be obtained using molecular epidemiological tools (Leitner and Albert 1999; Korber, Muldoon et al. 2000; Lewis, Hughes et al. 2008; Hué, Gifford et al. 2009; Sharp and Hahn 2011). The principle of phylogenetic analysis of genetic sequences is that as replication and propagation of the genetic information occurs, mutation and evolution generates genetic differences in the progeny of originally highly similar sequences that increase over time with ever-increasing rounds of replication and propagation of the genetic material. This assumption is the basic principle behind reconstruction of phylogenetic trees, and at its heart is the concept that the further away two genetic sequences are from their most recent common ancestor, the more genetic differences will have accrued between them. If the assumption of a molecular clock is made then the number of differences between two sequences is treated as directly proportional to the time since the two sequences diverged (Ho 2008). This assumption encounters a numbers of issues when faced with real data: firstly, the rates of nucleotide evolution across sequences are not equal, which is not surprising given different regions of a nucleotide sequence may be under different evolutionary pressures (Yang 1996; Huelsenbeck, Larget et al. 2000). In addition, the approach of simply counting differences between two sequences is confounded by the redundant nature of the genetic code (Yang 1994; Yang and Rannala 2012); for example, the observed distance between two nucleotide sequences may be at a single position, where there is an A in one sequence, and a C in the other, so there initially appears to be a single substitution event that has taken place between the two sequences. However, the possibility that the second sequence initially underwent

substitution to a G or T at that location, and subsequently a second substitution to a C cannot be ruled out, but cannot be directly discovered by observation of the final sequences. If the second sequence underwent an additional substitution back to an A, then the sequences would appear to have remained identical, despite the second sequence having actually gone through three substitution events. To address these issues a variety of models of nucleotide evolution have been developed to try to more realistically model the process of substitution (Yang and Rannala 2012). These models are based upon the principle of Markov chains, where the next step in the chain only depends upon the current state of the chain, and is not influenced by the state of the chain prior to the current state. Transition matrices are set up with differing probabilities attending to different substitutions depending on the model being used. Such probabilities may reflect differences in the likelihood of a substitution being a transition versus a transversion.

There are several approaches to constructing phylogenetic trees, broadly falling into i) those that rely upon distance measures between sequences, involving conversion sequences into matrices of genetic distances, and ii) those termed discrete methods, which act upon the sequences themselves (Whelan, Liò et al. 2001; Yang and Rannala 2012). Of the first approach, the neighbor-joining method is widely used. This distance-based method starts off with a completely unresolved tree, and then proceeds to cluster the most closely related sequences together first, using a matrix of computed distances, which are then replaced by their common ancestral node. The process continues to cluster pairs of taxa and recompute the overall distance matrix of the tree until the whole tree has been reconstructed. This approach is fast, but can be inaccurate, as it depends upon the order in which branches were added to the tree, and in the end the final tree that is computed cannot be compared to other trees that may better represent the underlying genetic relatedness of the sequences. This latter issue is a strength of the discrete method maximum-likelihood approach, which tries to

explore as many different tree topologies as possible, in order to find the tree that explains the sequence data best (i.e. the tree with the highest likelihood given the data). For this reason, maximum-likelihood approaches can identify much more accurate tree topologies, but require greater computational effort, especially if a large number of sequences are involved (Whelan, Liò et al. 2001; Yang and Rannala 2012). Both neighbor-joining and maximum-likelihood methods produce a single tree at the end of the process, which contains no information on the confidence of any particular sequence relationship within the tree. This information is typically obtained using non-parametric bootstrapping. This approach randomly samples, with replacement, columns from the original alignment in order to build up a new alignment that contains a subsample of the genetic information contained in the original sequence alignment (Felsenstein 1985). 100 or 1000 bootstrap replicates are used, depending on the size of the alignment, and new trees are constructed, ideally using the same process as the original tree-construction if computationally feasible. Support for a particular grouping of sequences within the original tree is obtained by observing the frequency at which this grouping appears in the bootstrap tree.

An additional approach to phylogeny reconstruction is to use a Bayesian statistical approach combined with a Markov Chain Monte Carlo algorithm (Yang and Rannala 1997; Whelan, Liò et al. 2001) to generate a sample of trees from which an overall consensus tree can be constructed. Each iteration of the algorithm proposes a new tree and set of parameters by altering the current tree and model parameters following a set of rules. The new tree may be accepted or rejected on the basis of particular criteria specified within the algorithm. A tree and parameter space of accepted proposals is built up until it is judged that the chain has generated a sufficient sample of trees, which can result in very long chain runs being required. The confidence in any one tree, or a cluster of sequences within a tree, is based upon the number of times it was proposed during the chain (Yang and Rannala 2012).

1.6.2 Practical issues: alignments, programs, trees

In terms of practical issues surrounding the construction of phylogenetic trees, it is crucial to use a sequence alignment of good quality, particularly if the sequences are likely to have regions containing gaps, such as in alignments of the variable loops in the HIV-1 envelope region. A number of alignment algorithms and tools exist, such as ClustalW and MUSCLE, but manual curation of alignments is also advisable where feasible, prior to inputting sequences into phylogenetic reconstruction tools (Yang and Rannala 2012).

A number of tree construction algorithms specialised for extremely large alignments have been developed, such as RAxML and FastTree. These methods use a variety of approximations and heuristic approaches to reduce the overall computational effort involved in tree construction, whilst attempting to maintain accuracy and reliability (Price, Dehal et al. 2010; Yang and Rannala 2012; Stamatakis 2014).

On a further practical note, it is clear that in terms of HIV-1, a virus prone to recombination *in vivo*, and where co-infection of individuals with highly unrelated viral subtypes can occur, that very different phylogenies may be obtained depending on the region of the virus used (Robertson, Sharp et al. 1995). There has been some debate as to the suitability of different regions of the HIV-1 genome to reconstruct epidemiological relationships within the epidemic (Hué, Clewley et al. 2004; Stürmer, Preiser et al. 2004; Hué, Clewley et al. 2005). The *pol* region has been at the centre of this discussion by virtue of the fact that the region is sequenced as part of routine drug resistance surveillance, and as such, there are a large number of *pol* sequences available in a number of countries where such surveillance is widely implemented, presenting a potentially highly useful resource for reconstruction and investigation of local or national epidemics (Lewis, Hughes et al. 2008; Yerly, Junier et al. 2009; Brenner, Roger et al. 2011; Foster, Ambrose et al. 2014). However, sequencing technologies, such as Single Molecule Real-Time (SMRT) sequencing, may

enable more routine whole genome sequencing in the near future, which may make such controversies a moot point going forward (Henn, Boutwell et al. 2012; Brown, Guo et al. 2013).

1.6.3 Defining transmission clusters

A number of studies have supported the ability of molecular epidemiological techniques to accurately reconstruct actual epidemiological events, where such data has been available from cases of known infection events (Patient 1992; Hillis and Huelsenbeck 1994; Leitner and Albert 1999; Paraskevis, Magiorkinis et al. 2004; Brooks, Robbins et al. 2006; Scaduto, Brown et al. 2010; Vandamme and Pybus 2013). Key to this is the identification of clusters of highly related viral genomes relative to surrounding genomes, and therefore the criteria used to define clusters highly likely to represent networks of direct infection. It is necessary to bear in mind what questions are being asked when identifying such clusters, as all HIV-1 will have been transmitted by a donor, and to that extent a most common recent ancestor (MCRA) will be identifiable for all sequences on a tree (Volz, Koopman et al. 2012). However, identification of transmission clusters is most often in the context of analysing aspects of the HIV-1 epidemic that may play a significant role in the onward spread or increased pathogenesis of the disease, such as the role of individuals with recent infection, or the impact of individuals who have failed treatment and may harbour drug resistant strains of the virus (Brenner, Roger et al. 2007; Lewis, Hughes et al. 2008; Yerly, Junier et al. 2009). In these contexts it is common practice to employ some measure of cluster support, e.g. bootstrap or posterior probability, together with an intra-cluster genetic distance cut-off. There is no agreed combination of support and distance cut-off used to define clusters, and studies vary depending on the questions being asked (Kaye, Chibo et al. 2008; Bezemer, van Sighem et al. 2010; Chalmet, Staelens et al. 2010). Necessarily, altering the genetic distance component of the cluster identification process will result in greater or lesser average

pairwise genetic distances between individuals within those clusters, and may influence downstream inferences based upon belonging to those clusters (Volz, Koopman et al. 2012). It also important to be aware that phylogenies of infected individuals are unlikely to be complete, given the numbers of individuals infected with HIV-1 that are unaware of their infection, and who will therefore not have viral sequence available for analysis (Campsmith, Rhodes et al. 2010; Lodwick, Alioum et al. 2011; Aghaizu, Brown et al. 2013).

1.7 Modelling of HIV-1 infection dynamics in populations

1.7.1 Mathematical models of transmission

In order to gain insight into the HIV-1 epidemic in various settings and populations, researchers have attempted to use mathematical models to answer questions relevant to the design and implementation of public health strategies to try to reduce the growth of the epidemic, such as the infectiousness of particular stages of HIV-1 infection or the percentage of individuals with undiagnosed infections (Pinkerton 2007; Hollingsworth, Anderson et al. 2008; Granich, Gilks et al. 2009; Campsmith, Rhodes et al. 2010). Despite the high viral loads associated with early HIV-1 infection, modelling of the probability of transmission per coital act and assessment of the contribution from early infection has produced different results through use of different models and parameters (Miller, Rosenberg et al. 2010). This is partly to do with the difficulty of obtaining accurate data on the rate of transmission from individuals in the acute infection to uninfected individuals who then go on to seroconvert. Several studies have used data obtained as part of a community randomised trial of sexually transmitted disease (STD) control for the prevention of AIDS, in Rakai District, Uganda (Wawer, Gray et al. 1998; Wawer, Sewankambo et al. 1999; Wawer, Gray et al. 2005; Abu-Raddad and Longini Jr 2008; Hollingsworth, Anderson et al. 2008). This study captured data

on couples where both partners were seronegative, and then one partner became seropositive, followed some time later by the second partner. Different treatment of the same data produced broadly similar agreement in their findings of increased infectiousness of the early and late phases of the disease, but differed in terms of the estimates of the infectiousness and the length of those periods, and their overall contribution to onward transmission within an epidemic, given that the asymptomatic period, with its lower rate of transmission, can often last for many years (Wawer, Gray et al. 2005; Hollingsworth, Anderson et al. 2008). A number of studies support the idea that the phase of the overall epidemic impacts upon the contribution from each phase of the disease, with early epidemics being largely driven by individuals in the early phases of infection, by virtue of the fact that insufficient time has passed for substantial numbers of individuals to have passed into the later phases of the infection (Abu-Raddad and Longini Jr 2008; Hollingsworth, Anderson et al. 2008). These and other studies also point out that other aspects of an epidemic, such as the population and the sexual risk behaviour of individuals, are likely to be important (Jacquez, Koopman et al. 1994; Koopman, Jacquez et al. 1997; Abu-Raddad and Longini Jr 2008).

1.7.2 Introduction to phylogenetic based models

Phylogenetic approaches have also been employed to examine transmission dynamics at play within epidemics. Using clusters of transmissions to identify particular risk groups disproportionately driving onward transmission has tended to identify the recent infection period of infection as playing a key role (Yerly, Vora et al. 2001; Pao, Fisher et al. 2005; Brenner, Roger et al. 2007; Fisher, Pao et al. 2010). However, the findings of these studies must be considered with caution, as sampling of an infected population is often incomplete, and linkage of those individuals who are sampled, and whose viruses were obtained from the recent phase of infection, does not mean that those individuals actually transmitted from the recent phase of infection (Brown, Gifford et al. 2009). On top of this, phylogenetic tree

construction methods that incorporate genetic distance measures may introduce inherent bias into the process, which makes the clustering of more highly related sequences more likely (Volz, Koopman et al. 2012). Chapter four of this thesis investigates the utility of a phylogenetic approach in combination with a marker of infection length to gauge the role of individuals with recent HIV-1 infection in driving the UK epidemic.

1.8 Ultra-deep sequencing of early HIV-1

1.8.1 Deep-sequencing theory: current methodologies and limitations

Until relatively recently, Sanger sequencing was the predominant way in which genetic information was generated for analysis (Sanger, Nicklen et al. 1977; Lander, Linton et al. 2001). However, there are now a variety of technologies available, often generically referred to as 'next-generation' sequencing technologies, that enable the production of genetic information on a large scale, at ever-decreasing costs (Metzker 2009; Koboldt, Steinberg et al. 2013). Two main platforms, Illumina and 454, both rely upon clonal amplification of template DNA prior to sequencing based upon measuring incorporation of nucleotides using fluorescence. The clonal amplification from HIV-1 samples allows a much greater spectrum of the quasi-species to be captured, and therefore enables a greater depth of sampling than traditionally possible with Sanger sequencing, which in effect generates a consensus sequence from the quasi-species by collapsing the population sub-structure into an overall 'average' sequence. This 'next-generation' technology therefore allows exploration of topics such as quasi-species dynamics in the presence of host immune pressure, and the impact of low level drug resistance on antiretroviral treatment outcomes.

1.8.2 454 methodology outline: tagged amplicons, MIDs

To date, the Illumina and 454 platforms have been the main technologies offering deep-sequencing capability useful for studying viruses. Through differences in technology, Illumina can produce a greater number of reads from a sample at a lower cost than 454 pyrosequencing (Glenn 2011), but generates shorter reads, in the region of 100-150bp in length, whereas 454 pyrosequencing can generate read lengths of 400bp and possibly longer. These differences in output lend themselves to different applications, and pyrosequencing has generally been useful for sequencing amplicons from specific regions of the HIV-1 genome, the V3 loop region of HIV-1 for example, or to reconstruct subpopulations, which requires reads to have sufficient overlap with each other to enable confident contiguation (Zagordi, Bhattacharya et al. 2011; Zagordi, Däumer et al. 2012). Briefly, the principle behind 454 pyrosequencing relies on an initial clonal amplification of individual molecules of DNA derived either unamplified from the original host if sufficient material is available, or from PCR. The clonal amplification is carried out in such a way that one individual molecule of DNA is clonally amplified in its own emulsion PCR reaction, which takes place on a picotitre plate, such that one emulsion PCR reaction of one DNA molecule takes place per well of the plate. The actual sequencing step is subsequently performed by breaking the emulsion PCR microreactor (water droplets within an oil phase) and adding additional beads with sulphurylase and luciferase attached. This step is followed by flowing a particular nucleotide across the plate and detecting incorporation events in particular wells where the nucleotide joins the complementary strand of a template DNA molecule. Incorporation of a nucleotide releases a pyrophosphate, which is detected by a high-resolution camera, thereby identifying every well where a nucleotide was incorporated (or more than one nucleotide in the case of homopolymeric regions in the template DNA) (Metzker 2009).

Genetic barcode systems for both Illumina and 454 allow multiple patients to be uniquely tagged with a short nucleotide sequence, Multiplex IDentifier (MID) tags, which means that all reads generated from that sample can be identified in downstream analyses and linked back to the patient sample, which can increase efficiency and reduce costs by decreasing the number of runs required to sequence large numbers of patients in parallel (454 Life Sciences Corp. 2009; Ji, Li et al. 2011; Dudley, Chin et al. 2012).

There are a number of limitations and potential biases that are associated with the different deep-sequencing technologies, and it is important to be aware of these during template preparation and downstream analysis steps. A major potential source of error during deep-sequencing is the generation of PCR errors during the sample preparation step (Hughes and Totten 2003; Kanagawa 2003; Jabara, Jones et al. 2011). The Taq polymerase reaction can incorporate 1.8×10^{-4} - 8.0×10^{-6} nucleotide errors per base pair per replication cycle (Terpe 2013), and can therefore introduce spurious low level variants into the results that will inflate estimates of true variants. The process of pyrosequencing itself encounters issues when dealing with homopolymer stretches of nucleotides, and can artificially elongate such regions (Huse, Huber et al. 2007). Though not specific to pyrosequencing, there is also the potential issue of low level swapping of MID tags during preparation of the sample for pyrosequencing, and it is necessary to take this into account when deep-sequencing in the context of viral quasi-species, where samples can potentially contain highly related viruses through transmission events, and/or can come from the same patient at multiple different time points (Carlsen, Aas et al. 2012). Knowledge of these issues, and strategies to ameliorate or correct for them, are essential in studies looking at low level variants and quasi-species diversity, in order to fully harness the power that these deep-sequencing technologies offer (Rozera, Abbate et al. 2009; Archer, Rambaut et al. 2010; Hedskog, Mild et al. 2010; Jabara, Jones et al. 2011; Zagordi, Bhattacharya et al. 2011).

Research aims

1. Develop a measure of length of HIV-1 infection based upon the analysis of the proportion of ambiguous nucleotides in Sanger *pol* consensus sequences (detailed in chapter three).
2. Construct phylogenies of Sanger *pol* consensus sequences from the UK HIV Drug Resistance Database to detect clusters of transmission, with the aim of identifying the extent to which individuals with recent infection are contributing to the spread of HIV-1 in the UK and their contribution to trends in transmitted drug resistance (detailed in chapter four).
3. Perform deep-sequencing of plasma HIV-1 RNA sampled from UK patients diagnosed with recent infection, in order to assess the complexity and evolution of viral quasi-species, and investigate the extent to which early infections may be founded by diverse viral subpopulations (detailed in chapter five).
4. Investigate the molecular epidemiology of HIV-1 infection in the Kumasi region of Ghana among newly diagnosed patients attending the Komfo Anokye Teaching Hospital, in order to identify the circulating subtypes and detect transmitted drug resistance (detailed in chapter six).
5. Perform phylogenetic analyses of HIV-1 infections in the Kumasi cohort to explore the extent to which clusters of highly related infections exist, which may indicate high risk behaviours that drive the local epidemic (detailed in chapter six).
6. Investigate the performance of methods for identifying recent HIV-1 infection (guanidine based avidity assay and the nucleotide ambiguity classifier) in the Kumasi cohort, with a view to investigating the utility of such approaches in gauging the role of individuals with recent (detailed in chapter six).

Chapter Two

Materials and methods

2 Materials and methods

2.1 Patient cohorts

2.1.1 UK HIV Drug Resistance Database

The UK HIV Drug Resistance Database (UKHIVDRD) (http://www.hivrdb.org.uk/) is a national repository of protease and reverse transcriptase sequences obtained by Sanger sequencing of HIV-1 obtained from patients undergoing drug resistance testing as part of routine care. At the time of the analysis, there were 43002 anonymised patients in the database, and a total of 55556 sequences from both antiretroviral treatment (ART)-naïve and ART-experienced patients. Associated clinical data for 16362 database patients was provided by the UK Collaborative HIV Cohort (http://www.ukchic.org.uk/). Sequences were supplied pre-aligned individually to a reference sequence (consensus B) using the LAP alignment program (Huang and Zhang 1996).

2.1.2 UK-based cohorts for the development of the nucleotide ambiguity cut-off

Viral sequences and clinical data were obtained from three cohorts: i) patients that attended the Ian Charleson Day Centre of the Royal Free Hospital (RFH), London, between 2004 and 2010; 2) 670 patients that between 1997 and 2009 became part of the UK Register of HIV Seroconverters, a UK-wide database of newly diagnosed patients whose time of seroconversion can be reliably estimated using laboratory evidence, and; 3) patients that attended the HIV clinic of St Mary's Hospital (SMH), London, between 2008 and 2010. Clinical data from all three cohorts was stripped of patient identifiable information prior to being made available to the study.

2.1.3 Kumasi HIV new diagnosis cohort

Serum and plasma samples were collected from randomly selected newly diagnosed patients attending the HIV clinic at Komfo Anokye Teaching Hospital in Kumasi, Ghana, between

2008 and 2012. Associated clinical and laboratory data were collected from the case records and anonymised. The study was approved by the Committee on Human Research Publications and Ethics at Kwame Nkrumah University of Science and Technology (KNUST), Kumasi, Ghana.

2.2 Reference sequences

HXB2 was used as a reference genome for all primer and sequence coordinates (Genbank Accession number K03455).

2.3 HIV-1 Avidity assay

Guanidine-based avidity assays were carried out using the Vitros ECiQ Immunodiagnostic System (Ortho-Clinical Diagnostics, United Kingdom) as per the Operator's Guide. The machine was calibrated using the Vitros Anti-HIV 1+2 Calibrator when appropriate (e.g. when reagent pack lot number changed, and/or at least once every 28 days upon change of reagent pack and calibrator lot). Before each avidity run, a quality control step was performed using the Vitros Anti-HIV 1+2 Controls. Control sera were reconstituted in 1ml of de-ionized water. Additional high and low avidity control samples were included at the start and end of each run. Each sample (controls and patient samples) were run in duplicate following incubation for 10 min at a 1:10 (20µl sample in 180µl diluting solution) dilution in either phosphate-buffered saline (reference dilution) or 1M Guanidine Hydrochloride (test dilution). The avidity index (AI) was calculated by dividing the sample-to-cut-off ratio of the test dilution (mean of replicate wells) over the sample-to-cut-off ratio of the reference dilution (mean of replicate wells). An avidity index of ≤ 0.75 was taken to be indicative of the patient

having undergone HIV-1 seroconversion within the previous 125 days (95% confidence interval [CI], 85 to 164 days).

2.4 Nucleic acid extraction using the NucliSENS easyMAG extraction platform

Plasma and serum samples were extracted using the NucliSENS easyMAG (bioMerieux, Basingstoke, UK) for total nucleic acid extraction platform. Frozen plasma/serum samples stored at -80°C were brought to room temperature and vortexed for >10s to ensure homogenization. Depending on viral load of the sample, and the volume of material available, 0.2ml, 0.5ml or 1.0ml of each sample was loaded onto the NucliSENS easyMAG and total nucleic acid was extracted as per manufacturer guidelines, using an onboard lysis step and Generic 2.0.1 protocol. A negative water control was included in each extraction run. Elution volume was 25µl, 35µl or 65µl of easyMag elution buffer 3 depending on volume required for downstream applications. Eluate was transferred to ice immediately after extraction, prior to use in downstream applications. Any eluate remaining after downstream applications had been carried out was stored at -80°C.

2.5 One-step Reverse Transcription PCR (RT-PCR) reaction for *pol* amplification prior to nested PCR

HIV-1 *pol* in-house PCR primers (previously developed by colleagues at the Royal Free Hospital, London, as part of a diagnostic protocol):

Primer (HXB2 coordinates)	**Sequence**
RES1 (1819-1844bp):	GAA GAA ATG ATG ACA GCA TGT CAG GG
RES2 (4202-4173bp):	TAA TTT ATC TAC TTG TTC ATT TCC TCC AAT

RES3 (3585-3559bp):	ATG GYT CTT GAT AAA TTT GAT ATG TCC
RES4 (2074-2095bp):	AGA CAG GCT AAT TTT TTA GGG A

A master mix containing the following amounts of each reagent from the Qiagen OneStep RT-PCR Kit (Qiagen Cat. No. 210212) kit was prepared and 40µl aliquoted into each well of a 96-well plate:

Reagent	Amount **per sample** (µl)
5x buffer	10.0
dNTPs (10mM)	2.0
Primer RES1 (10µM)	3.0
Primer RES2 (10µM)	3.0
RT-PCR enzyme mix	2.0
Water	20

10µl of RNA was added to each well of the 96-well plate, together with appropriate negative controls, and the plate was transferred to a Veriti thermocycler, where the following program was initiated:

Temperature (^{o}C)	**Time**	**Cycles**
50	35 mins	1
95	15 mins	1
95	30 secs	1
65	45 secs	
72	3 mins	
95	30 secs	1
60	45 secs	

72	3 mins	
95	30 secs	45
58	45 secs	
72	3 mins	
72	10 mins	1
4	Hold (<24 hours)	1

2.6 Nested PCR for *pol* amplification prior to cycle-sequencing

Upon termination of the RT-PCR, a master mix of the following reagents from the Qiagen HotStarTaq DNA Polymerase (Qiagen Cat. No. 203205) kit was prepared, and 95µl aliquoted into the relevant wells of a 96-well plate:

Reagent	**Amount per sample (µl)**
Buffer (10x)	10.0
dNTPs (10mM)	2.0
Primer RES3 (10µM, note 1:10 dilution of stock)	2.0
Primer RES4 (10µM, note 1:10 dilution of stock)	2.0
Hot start Taq (5U/µl)	0.5
Water	78.5
Total	*95*

5µl of RT-PCR product was added to the relevant wells and the plate transferred to a Veriti thermocycler, where the following program was run:

Temperature (^{o}C)	**Time**	**Cycles**
95	12 mins	1
95	30 secs	1

65	45 secs	
72	3 mins	
95	30 secs	1
60	45 secs	
72	3 mins	
95	30 secs	45
55	45 secs	
72	3 mins	
72°C	10 mins	1
4°C	Hold (<24 hours)	1

Upon termination of the nested PCR, visualisation of the DNA amplicons via gel electrophoresis was carried out using a 2% agarose gel with SYBR Safe DNA Gel Stain (Invitrogen Cat. No. S33102) and HyperLadder 1kb, formerly HyperLadder I (Bioline Cat. No. BIO-33053), with an expected sample amplicon size of 1511bp. A band in the negative control was taken to be indicative of PCR contamination.

PCR products of the correct size were purified prior to cycle-sequencing using the QIAquick PCR Purification Kit (Qiagen Cat. No. 28106), as per manufacturer instructions. 65µl of molecular biology grade water was then added to each sample to ensure sufficient volume of material for cycle-sequencing with each of the 6 primers (see 2.8).

2.7 Reverse transcription reaction for *env* prior to PCR and Sanger sequencing

A master mix of the following amounts of each reagent from the Invitrogen SuperScript III First-Strand Synthesis SuperMix kit (Invitrogen Cat. No. 18080-400) together with reverse primer 16R (5-GGT AGC TGA AGA GGC ACA GG-3) was prepared on ice. Depending on

the viral load of a sample, and the downstream application, half or third volumes were used, in order to conserve reagents, with corresponding alterations in eluate input volume. 5µl of master mix was aliquoted into separate wells of a 96-well plate and kept on ice.

Reagent	**Amount per sample (µl)**
Annealing buffer	2.5
Molecular biology grade water	1.87
16R (20uM)	0.63
Total	*5*

15µl of sample eluate was transferred into the corresponding well of the 96-well plate (including an additional well for the negative water control extract and a positive control where appropriate) and the plate transferred to a Veriti thermocycler, where the following program was initiated in order to achieve template-primer annealing:

Temperature (°C)	**Time**	**Cycles**
65	5 mins	1

At termination of the program, the plate was immediately transferred to ice for 1 minute, before the following enzyme master mix was added to each sample well:

Reagent	**Amount per sample (µl)**
SuperScript® III/RNaseOUT™ Enzyme Mix	5
2X First-Strand Reaction Mix	25
Total	*30*

The plate was covered with an adhesive seal and transferred to a Veriti machine on ice, and the following thermocycler program initiated:

Temperature (°C)	Time	Cycles
55	50 mins	1
85	5 mins	1
4	∞	1

Once the thermocycler program has terminated, cDNA was stored at 4°C until required for the polymerase chain reaction amplification step. Long term storage of any remaining cDNA was at -20°C/-80°C.

2.8 Polymerase chain reaction amplification of *env* prior to Sanger sequencing

***Env* PCR primer sequences:**

Primer (HXB2 coordinates)	Sequence
2F (5559-5577bp)	ATG GAA CAA GCC CCA GAA G
4F (5966-5984bp)	TCC TAT GGC AGG AAG AAG C
15R (8365-8344bp)	GGT GAG TAT CCC TGC CTA ACT C
16R (8529-8510bp)	GGT AGC TGA AGA GGC ACA GG

A master mix containing the following amounts of each reagent from the Invitrogen Platinum PCR SuperMix High Fidelity (Invitrogen Cat. No. 12532-024) was prepared per sample, and 48µl aliquoted into a 96-well plate. Depending on the viral load of a sample, and the downstream application, half or third volumes were used, in order to conserve reagents, with corresponding alterations in cDNA input volume:

Reagent	Amount per sample (µl)
Platinum *Taq* DNA Polymerase High Fidelity	45

2F (10uM)	1.25
16R (10uM)	1.25
Molecular biology grade water	0.5
Total	*48*

2µl of cDNA was added to corresponding wells of the 96-well plate, together with appropriate positive and negative controls.

The plate was covered with an adhesive seal and transferred to a Veriti thermocycler heating block, whereupon the following program was initiated:

Temperature (°C)	**Time**	**Cycles**
94	2 mins	1
94	15 secs	30
50	30 secs	
68	4 mins	
68	10 mins	1
4	∞	1

Upon termination of the first round PCR, a nested PCR master mix containing the following amounts of each reagent from the Invitrogen Platinum PCR SuperMix High Fidelity (Invitrogen Cat. No. 12532-024) was prepared per sample, and 48µl aliquoted into a 96-well plate. Depending on the viral load of a sample, and the downstream application, half or third volumes were used, in order to conserve reagents, with corresponding alterations in DNA input volume:

Reagent	**Amount per sample (µl)**
Platinum *Taq* DNA Polymerase High Fidelity	45
4F (10uM)	1.25
15R (10uM)	1.25
Molecular biology grade water	0.5
Total	*48*

In a designated post-amplification ultra-violet (UV) light cabinet, 2µl of PCR product from the first round plate was transferred to the corresponding well of the nested PCR plate using a 10µl multi-channel pipette, and the plate covered with an adhesive seal and transferred to a Veriti thermocycler heating block, whereupon the following program was initiated (the first round PCR plate was stored at -20°C/-80°C freezer for long term storage if necessary):

Temperature (°C)	**Time**	**Cycles**
94	2 mins	1
94	15secs	35
50	30 secs	
68	3.5 mins	
68	10 mins	1
4	∞	1

Upon termination of the nested PCR, visualisation of the DNA amplicons via gel electrophoresis was carried out using a 2% agarose gel with SYBR Safe DNA Gel Stain (Invitrogen Cat. No. S33102) and HyperLadder 1kb, formerly HyperLadder I (Bioline Cat. No. BIO-33053), with an expected sample amplicon size of 2399bp. A band in the negative control was taken to be indicative of PCR contamination.

PCR products of the correct size were purified prior to cycle-sequencing using the QIAquick PCR Purification Kit (Qiagen Cat. No. 28106), as per manufacturer instructions. 65µl of molecular biology grade water was then added to each sample to ensure sufficient volume of material for cycle-sequencing with each of the 12 primers (see below).

2.9 Cycle sequencing and clean-up

HIV-1 *pol* sequencing primers used in cycle-sequencing:

Primer (HXB2 coordinates)	**Sequence**
SEQ1 (2149-2165bp):	GAG CCA ACA GCC CCA CC
SEQ2 (2616-2634bp):	CAA TGG CCA TTG ACA GAA G
SEQ3 (3012-3031bp):	GGA TCA CCA GCA ATA TTC CA
SEQ5 (2606-2586bp):	TGG GCC ATC CAT TCC TGG CTT
SEQ6 (2988-3007bp):	CAT CCC TGT GGA AGC ACA TT
SEQ7_3 (3549-3569bp):	TTG ATA TGT CCA TTG GCC TTG

HIV-1 *env* sequencing primers used in cycle-sequencing:

Primer (HXB2 coordinates)	**Sequence**
4F (5966-5984bp)	TCC TAT GGC AGG AAG AAG C
6F (6340-6359bp)	ATT ATG GGG TAC CTG TGT GG
ENV2 (6561-6580bp)	GAT CAA AGC CTA AAG CCA TG
6955FQ (6955-6973bp)	CAG TAC AAT GYA CAC ATG G
OFM54 (7350-7374bp)	TTT AAT TGT GGA GGG GAA TTT TTC T
TUG (7859-7879bp)	GTC TGG TAT AGT GCA ACA GCA
6582R (6582-6558bp)	CAC ATG GCT TTA GGC TTT GAT CCC A
TUE3 (7006-7028bp)	TCC TTC TGC TAG ACT GCC ATT TA
V3-2M (7371-7352bp)	AAA ATT CCC CTC CAC AAT TA
EDS8 (7648-7668bp)	CAC TTC TCC AAT TGT CCC TCA

14R (8037-8017bp)	TGC AGA TGA GTT TTC CAG AGC
15R (8365-8344bp)	GGT GAG TAT CCC TGC CTA ACT C

For cycle-sequencing of either *pol* or *env*, the cycle-sequencing master mix was prepared with the following reagents per sequencing primer per sample, including BigDye ready reaction mix from the BigDye Terminator v3.1 Cycle Sequencing Kit (Invitrogen Cat. No. 4337456):

Reagent	**Amount per sample (µl)**
BigDye ready reaction mix	2
Molecular biology grade water	9
Total	*12*

The master mix was transferred to a 96-well plate and 8µl of each sample was added to 6 or 12 wells depending on the region being sequenced. 1µl of each of the cycle-sequencing primers (10uM) was added to a separate well. Once all samples had been added to the corresponding wells, the plate was transferred to a Veriti thermocycler for cycle sequencing with the following conditions:

Temperature (°C)	Time	Cycles
96	10 secs	25
50	5 secs	
60	4 mins	
4	∞	1

Upon termination of the cycle-sequencing thermocycler program, a fresh solution of sodium acetate and ethanol was made up as follows:

	Per well	E.g. 96-well plate (x1.2)
Sodium Acetate 3.0M pH4.6	2	240
100% ethanol (high grade)	50	6000
Total	*52*	*6240*

52µl of the above solution was added to each well and gently aspirated 3-4 times to mix. The plate was then covered with an adhesive plate seal and centrifuged at 2000g for 20 minutes at room temperature.

The plate was then transferred from the centrifuge, the seal removed, and the plate inverted onto absorbent paper towels before centrifuging upside down at 150g for 1 minute at room temperature.

150µl of 70% ethanol (freshly made up that day) was then added to each well and the plate covered with an adhesive seal. The plate was then centrifuged right-way-up at 2000g for 5 minutes at room temperature.

The plate was then transferred from the centrifuge, the seal removed, and the plate inverted onto absorbent paper towels before centrifuging upside down at 150g for 2 minutes at room temperature. The plate was then allowed to air-dry right-way-up for at least 30 minutes to ensure complete evaporation of residual ethanol.

DNA pellets were then resuspended in 10-20uI of Hi-Di Formamide (Applied Biosystems Cat. No. 4311320) (if they were to be run on the 3730 Genetic Analyser that day) or covered with a plate sealer and stored at -20°C.

2.10 Sequence analysis in SeqScape and other tools

Products were analysed on an Applied Biosystems 3730 DNA Analyzer and manually curated using SeqScape v2.6 (Applied Biosystems, Warrington, United Kingdom).

2.11 454 deep-sequencing

All samples were normalised to a viral load of 20000 copies/ml using basematrix to ensure approximately identical virus input. Extraction was carried out using the NucliSENS easyMAG total nucleic acid extraction platform into an elution volume of 35µl (in NucliSENS easyMAG extraction buffer 3).

A clone control was incorporated into the experiment in order to ascertain levels of random errors incorporated into reads through a combination of PCR, emPCR and pyrosequencing errors (because the clone control was provided as DNA, it is not able to act as a control for the reverse transcription step). The control clone used was ARP2071 (pCRII-TOPO-SE8131-3) plasmid DNA (a full length clade A clone (Accession Number: AF107771) in a pCRII-TOPO vector), obtained from the National Institute for Biological Standards and Control (NIBSC) Centre for AIDS Reagents (http://www.nibsc.org/spotlight/centre_for_aids_reagents.aspx).

2.11.1 RT and PCR of *pol* and *env* region of HIV-1

A master mix was prepared on ice containing the following amounts of each reagent from the Invitrogen SuperScript® III First-Strand Synthesis SuperMix:

Reagent	**Amount per sample (µl)**
Annealing buffer	2.5
Molecular biology grade water	1.87
Reverse primer* (20uM)	0.63
Total	*5*

**pol* and *env* amplicons were generated using two different primer sets, with two different outer reverse primers for cDNA generation: JA272 (5-GGATAAATCTGACTTGCCCART-3) for *pol* and 15R (5-GGTGAGTATCCCTGCCTAACTC-3) for *env* (i.e. two separate master mixes were generated).

5µl of each master mix was aliquoted into a 96-well plate and covered with an adhesive plate seal. 15µl of each eluate was transferred into corresponding wells of the 96-well plate (i.e. 15µl into the *pol* well and 15µl into the *env* well for each sample), together with positive and negative controls. The plate was covered with an adhesive seal and transferred on ice to a Veriti thermocycler heat-block where the following program was run:

Temperature (°C)	Time	Cycles
65	5 mins	1

Upon termination of the program, the plate was removed to ice and left for 1 minute prior to the addition of 30µl of the following enzyme master mix per sample well:

Reagent	**Amount per sample (µl)**
SuperScript® III/RNaseOUT™ Enzyme Mix	5
2X First-Strand Reaction Mix	25
Total	*30*

The plate was then transferred to a Veriti thermocycler heat-block on ice, and the following program run:

Temperature (°C)	Time	Cycles
55	50 mins	1
85	5 mins	1
4	∞	1

The following PCR master mix was prepared with the following reagents from the Invitrogen Platinum® Taq DNA Polymerase High Fidelity kit:

Reagent	**Amount per sample (µl)**
10X High Fidelity PCR Buffer	5.5
10mM dNTP mixture	1.1
50mM MgSO4	2.2
Primer outer F‡ (25µM each)	0.44
Primer outer R‡ (25µM each)	0.44
Platinum® Taq High Fidelity	0.22
Autoclaved, distilled water	30.1
Total	*40*

‡*pol* and *env* require two separate master mixes, each with a different primer combination:

	Primer outer F	**Primer outer R**
Pol	JA269 (5-AGGAAGGACACCARATGAARGA-3)	JA272 (sequence above)
Env	4F (5-TCCTATGGCAGGAAGAAGC-3)	15R (sequence above)

At this point the ARP2071 clone control was incorporated as one of the samples, and treated identically to patient sample cDNA.

10µl of RT product from above was transferred to the corresponding well containing the PCR reaction mix, the plate sealed and transferred to a Veriti thermocycler heat-block where the following program was run:

Temperature (°C)	**Time**	**Cycles**
94	2 mins	1
94	15 secs	20
55	30s	
68	3 mins	
68	10 mins	1
4	∞	1

A nested PCR master mix was prepared using the following reagents per sample, and 49µl transferred to the wells of a 96-well plate:

Reagent	**Amount per sample (µl)**
Platinum® PCR SuperMix High Fidelity	45
Primer inner F[†] (25 µM each)	0.44
Primer inner R[†] (25 µM each)	0.44
Total	*49*

[†]these primers were region specific, but had GS FLX Titanium emPCR Kit LibA fusion adaptors attached to enable generation of amplicons that could be hybridized to the DNA Capture Beads prior to emulsion PCR:

	Primer inner F	**Primer inner R**
Pol	NG_PinF2	NG_PinR2
Env	6955FQ	NG_EinR2

16 different paired Multiplex Identifier Tags (MID) tags were used, to enable multiplexing of 16 samples per pool (Appendix 1).

1µl of first-round product was transferred into the nested PCR plate, the plate was sealed, and transferred to a Veriti thermocycler heat-block and the following program was run:

Temperature (°C)	**Time**	**Cycles**
94	2 mins	1
94	15 secs	30
48	30 secs	
68	3 mins	
68	10 mins	1
4	∞	1

After the nested PCR, amplicons were visualised on a 2% agarose gel, and products of the correct size taken forward for clean-up. The *pol* amplicon was expected to be 426bp in size, and the *env* amplicon was expected to be 421bp in size.

PCR products of the correct size were purified prior to cycle-sequencing using the QIAquick PCR Purification Kit (Qiagen Cat. No. 28106), as per manufacturer instructions.

2.11.2 Measuring DNA concentration using Qubit

Amplicons from different patient samples needed to be pooled at the same mass concentration (or molar concentration if the amplicons varied >50bp in size), which required accurate determination of DNA concentration in each sample using the Invitrogen Qubit® dsDNA HS Assay Kit, and the Qubit® 2.0 Fluorometer. This was carried out as per manufacturer guidelines. Samples were then diluted to a concentration of 10ng/µl, and pooled into non-duplicate MID tag pools

2.12 Margin of error calculations for minimum deep-sequencing subpopulation prevalence cut-off

Carried out as per the following formula: $\mathrm{E} = z * \sqrt{\frac{\acute{p}(1-\acute{p})}{n}}$

Where E is the margin of error, z is the confidence level (the value of 2.58 was used to obtain a 99% confidence level – meaning that if the sample was deep-sequenced 100 times, the number of reads obtained for the subpopulation in question should fall within the range calculated 99/100 times), ṕ is the proportion of reads in the sample corresponding to the subpopulation having its margin of error calculated, n is the sample size which corresponds to the total number of reads in the sample (Lohr 2009). The resulting number was then converted into the equivalent in percentage and subtracted from the percentage abundance of the subpopulation in question. If the new read percentage abundance then fell below 1%, the subpopulation was deemed unreliable and not taken forward for analysis. For example, if a putative subpopulation was present at 51 reads in 4581 total reads (i.e. 1.1%) the margin of error was calculated as 0.004, or 0.4%. Subtracting 0.4% from 1.1% would give a lower

boundary percentage of 0.7% (translating into 33 reads in 4581), which is below the 1% cut-off, and hence the subpopulation would be discarded.

2.13 Molecular epidemiological analysis

2.13.1 Alignment and trimming

Sequence alignments were generated using the MUSCLE (MUltiple Sequence Comparison by Log- Expectation) automated alignment program (Edgar 2004). Subsequent inspection of alignment results, manual editing and gap-stripping, was carried out using BioEdit version 7.1.3.0.

2.13.2 Counting nucleotide ambiguities in fasta sequences

Microsoft Excel was used to calculate the number of ambiguous nucleotides in each fasta sequence within the UKHIVDRD. The overall aligned sequence file was transferred to Microsoft Excel in a manner that ensured one sequence per row. All sequences were adjusted so that only uppercase letters were included (i.e. any lowercase letters were converted into uppercase). All gap symbols "-" were removed prior to counting ambiguities, and DRAM codons removed with reference to the WHO 2009 Drug Resistance Mutation List (Bennett, Camacho et al. 2009). The following formula was used to enumerate each IUPAC ambiguity symbol, where "$D1" is the cell reference to the gap-stripped, uppercase fasta sequence, and where "R" is the IUPAC ambiguity symbol:

```
=LEN($D1)-LEN(SUBSTITUTE($D1,"R",""))
```

Other IUPAC symbols used were Y,S,W,K,M,B,D,H and V. Whilst B,D,H and V represent mixtures of three bases, the presence of these symbols were counted as one nucleotide

ambiguity only. The ambiguity score was calculated as the sum of all IUPAC symbol counts per sequence divided by the length of the sequence, i.e. =LEN($D1).

2.13.3 Building trees (PhyML, BEAST, FastTree)

UKHIVDRD FastTree clusters

A UK HIV Drug Resistance Database sequence alignment was supplied by the Medical Research Clinical Trials Unit (MRC-CTU) which was generated using Local Alignment Program (LAP) to align sequences to a reference sequence (consensus B) (Huang and Zhang 1996). Subtype specific sequence alignments were taken forward to generate phylogenies using FastTree version 2.1 (Price, Dehal et al. 2010). A tree was inferred using the General Time Reversible model of nucleotide evolution option, together with the Gamma option to approximate the different rates of evolution across different sites. For each phylogeny generated, support for branches was assessed using the bootstrapping method. Phylip's SEQBOOT (http://evolution.genetics.washington.edu/phylip/doc/seqboot.html) was used to generate 1000 re-sampled alignments from the original alignment in question, and these 1000 bootstrap alignments were analysed in FastTree2.1 using the -n option to generate 1000 bootstrap trees. The Perl script CompareToBootstrap.pl (supplied on the FastTree website http://www.microbesonline.org/fasttree/) was used to compare the tree from the original alignment in question to the 1000 bootstrap trees.

2.14 In-house Perl scripts

2.14.1 Identifying transmission clusters

Identification of transmission clusters was carried out using a custom written Perl script utilising BioPerl and the Analyses of Phylogenetics and Evolution (APE) package in R

respectively (Paradis, Claude et al. 2004): findclusterdistances.pl (available from the author upon request), together with the Awk programming language and Linux command line (Appendix 2).

2.14.2 HIV-1 recent infection transmission models

Two transmission models described in results chapter 4 were written as custom Perl scripts: recentestablishedsorter_adapted_for_infector_assignment.pl together with genetic_distance_model.pl (Appendix 2), both available from the author upon request.

Chapter Three

Assessing the performance of a nucleotide ambiguity based measure of HIV-1 infection length

3 Assessing the performance of a nucleotide ambiguity based measure of HIV-1 infection length

Abstract

Individuals with recent HIV-1 infection are potentially major drivers of transmission cluster formation due to the high viral titres and ongoing risky sexual behaviour associated with this risk group. Methods exist to detect individuals in this category, in order to better understand the role they play in HIV-1 epidemics. Whilst the majority of these methods are laboratory based, this chapter develops and validates a method based upon detecting nucleotide ambiguities in *pol* consensus sequencing.

Using a stringently categorised cohort from the Royal Free Hospital, all with avidity assay results, a *pol* consensus sequence nucleotide ambiguity cut-off of 0.17% was identified as offering the best discrimination between recent and established infection in the development cohort (sensitivity and specificity of 85% and 82% respectively). A more stringent cut-off of 0.00% was also investigated. Validation on two additional cohorts showed a reduced ability to detect true recent infection using both cut-offs, though the 0.00% cut-off offered greater positive predictive power.

Overall, despite certain limitations, the approach may offer a useful way to identify individuals likely to have recent infection when only viral consensus sequence is available, such as in national HIV drug resistance databases.

3.1 Introduction

HIV epidemics in every setting, whether in a developing or developed country context, are driven by transmission from individuals at different stages of infection (Wawer, Gray et al. 2005; Boily, Baggaley et al. 2009). A variety of clinical and behavioural factors affect the likelihood of transmission from any particular stage, such as viral load (Pilcher, Tien et al. 2004), awareness of infection status (Marks, Crepaz et al. 2005) and antiretroviral treatment (ART) status (Donnell, Baeten et al. 2010). A number of studies support the hypothesis that the initial stage of HIV-1 infection plays a major role in onward transmission of the virus because of a combination of these factors. It is well established that acute infection is associated with high viral titres in plasma (Pilcher, Shugars et al. 2001; Fiebig, Wright et al. 2003), and correspondingly high viral titres in genital compartments (Pilcher, Joaki et al. 2007; Morrison, Demers et al. 2010). There is also evidence to support the idea that individuals recently infected with HIV-1 may be unaware of their infection status, and transmission rates from individuals unaware of their infection have been found to be higher than from diagnosed individuals (Marks, Crepaz et al. 2006). It is clearly important for policymakers determining national public health strategies to better understand their HIV epidemics. Defining the transmission dynamics of individuals in the initial stages of HIV infection, and the role that recent infection plays in onward transmissions is required to guide prevention measures designed to reduce onward spread of HIV.

To date, monitoring of the numbers of individuals in the initial stage of HIV infection has most commonly used laboratory assays that measure the evolving strength of the anti-HIV antibody response over time, such as guanidine-based avidity assays (Chawla, Murphy et al. 2007; Mastro, Kim et al. 2010) and branched peptide antigen (BED) based capture Enzyme-Immuno Assays (EIA) (Parekh, Kennedy et al. 2002). These are often then combined with other clinical biomarkers, $CD4^+$ cell count for example, into algorithms such

as the Serologic Testing Algorithm for Recent HIV Seroconversion (STARHS), developed by the US Centers for Disease Control and Prevention (CDC) (Janssen, Satten et al. 1998), or the Recent Infection Testing Algorithm (RITA) used by Public Health England (Garrett, Lattimore et al. 2012).

Recently (partly overlapping in time with the work carried out in this chapter), a number of studies have investigated using HIV-1 sequences obtained from patients as part of routine antiretroviral drug resistance surveillance to estimate infection length in individuals, without recourse to serological assays (Kouyos, von Wyl et al. 2011; Ragonnet-Cronin, Aris-Brosou et al. 2012; Andersson, Shao et al. 2013). This approach is based upon the assumption that the majority of HIV-1 infections are founded by a single virus (Zhang, MacKenzie et al. 1993; Zhu, Mo et al. 1993; Keele, Giorgi et al. 2008), which over time expands in a linear manner into a highly heterogeneous viral quasi-species that can be observed as mixed nucleotide peaks ("ambiguities") in electropherograms of consensus Sanger sequences. Kouyos et al. (2011) used three different cohorts of individuals, with their stage of infection estimated by different means (Bayesian back calculation using clinical information together with last known negative/first known positive dates; seroconversion estimated from last known negative/first known positive dates <180 days apart; and diagnosis of primary infection respectively) to generate a nucleotide ambiguity cut-off of >0.50% to indicate likelihood of an individual having been infected for >1 year. Ragonnet-Cronin et al. (2012) generated a training data set using individuals with p24 Ag+/Ab- results or negative test date within the 155 days prior to diagnosis as recently infected individuals, and patients with samples collected >155 days after diagnosis as established infections. The study also looked at whether concentrating on particular sequence sites might provide more information than using the overall sequence (e.g. sites of greater entropy, sites associated with HLA epitopes). They found that a cut-off of <0.45% nucleotide ambiguities best identified individuals with

recent infection (<155 days old), and that concentrating on sites that increased in entropy over the course of infection did improve sensitivity and specificity. The aim of this study was to combine the application of HIV antibody avidity testing and Sanger sequencing in a cohort of newly HIV-1 diagnosed patients with detailed clinical histories, in order to develop a nucleotide sequence ambiguity classifier that enables identification of recent HIV-1 infection. We assessed the utility of the proposed classifier in different cohorts and used deep-sequencing to explore the extent to which low level subpopulations may affect the utility of the approach.

3.2 Methods

3.2.1 Study population

Three anonymised cohorts of newly diagnosed HIV-1 infected patients were used to develop and validate the ambiguity classifier: 1) 517 patients that attended the Ian Charleson Day Centre of the Royal Free Hospital (RFH), London, between 2004 and 2010 ; 2) 670 patients that between 1997 and 2009 became part of the UK Register of HIV Seroconverters, a UK-wide database of newly diagnosed patients whose time of seroconversion can be reliably estimated using laboratory evidence, and; 3) 169 patients that attended the HIV clinic of St Mary's Hospital (SMH), London, between 2008 and 2010. Clinical data from all three cohorts was stripped of patient identifiable information prior to being made available to the study.

3.2.2 Recent infection testing

Guanidine-based avidity assay results were available for the Royal Free Hospital patient cohort through routine clinical care (assays were performed by diagnostic staff). The assay measures the strength of binding between immunoglobulin G antibodies and three HIV-1 antigens (p24, *env*10 and *env*13) using the anti-HIV-1 and -2 VITROS ECiQ assay (Ortho-Clinical Diagnostics, United Kingdom) (Chawla, Murphy et al. 2007). An avidity index of ≤0.75 reproducibly identifies seroconversion within the previous 125 days (95% confidence interval [CI], 85 to 164 days). A subset of the SMH cohort had results from a 'detuned' enzyme immunoassay (EIA) which identifies recently infected individuals using a modified protocol of the standard EIA used to detect HIV infection, by increasing the specimen dilution and decreasing the sample volume and substrate incubation times to make the assay less sensitive. A standard optical density (SOD) cut-off of ≤1.0 is associated with a HIV

infection of approximately 170 days (CI 95%: 163 to 183 days) (Rawal, Degula et al. 2003). The remainder of the SMH cohort (post-2008) had avidity results obtained using the AxSYM HIV 1/2 gO immunoassay (Abbott Diagnostics, Maidenhead, UK) (Suligoi, Galli et al. 2002).

3.2.3 Nucleotide sequence generation

Nucleic acid was extracted from plasma samples obtained from the RFH cohort, PCR amplified and cycle-sequenced as previously described (sections 2.4, 2.5 and 2.8) by members of the diagnostic staff at the Royal Free Hospital as part of routine diagnostic care. Ambiguous nucleotide peaks (i.e. mixed peaks, Figure 3.1) were identified using the default SeqScape mixed base peak cut-off of 25% (i.e., if a secondary peak reaches ≥25% of the height of the major peak, it is called as a mixed base), together with manual inspection of sequence electropherograms by individual members of the diagnostic staff at the Royal Free. *Pol* sequences were also retrieved from the UK Register of HIV Seroconverters, which receives sequence data generated by contributing laboratories across the UK using the Viroseq assay (Abbott, Maidenhead, UK), the TRUGENE genotyping kit (Bayer HealthCare, Berkeley, CA) or in-house methods. Finally, *pol* sequences obtained from the SMH cohort were generated by the Retrovirology Laboratory of Imperial College, London, using an in-house nested PCR approach employing primers located within conserved regions of the Gag and RT genes.

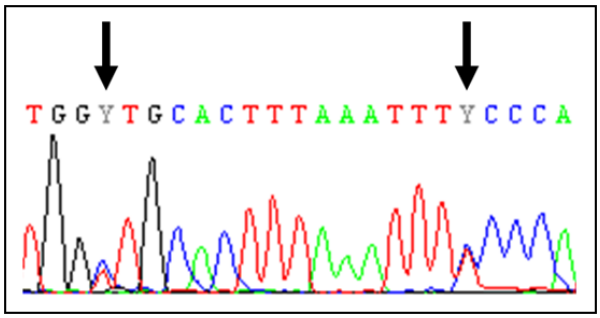


Figure 3.1. Screenshot of an electropherogram displayed in the SeqScape software, with arrows indicating mixed base peaks and with the IUPAC representative code above (Y for pYrimidine, i.e. C or T).

3.2.4 Nucleotide ambiguity analysis

All patient sequences were supplied as fasta files, and were stripped of codons associated with transmitted drug resistance mutations prior to subsequent nucleotide ambiguity analysis (Bennett, Camacho et al. 2009). International Union of Pure and Applied Chemistry (IUPAC) nucleotide ambiguity coded nucleotides were enumerated for each patient *pol* sequence, regardless of whether the code represented a bi-or tri- nucleotide mixture (e.g. 'R' and 'V' were both counted as single ambiguous nucleotides). 'N' IUPAC codes were not included in the ambiguous nucleotide count on the basis that they were likely to be indicative of poor quality sequence, rather than a tetra-nucleotide mixture.

3.2.5 Establishment of a nucleotide ambiguity cut-off

The 517 newly diagnosed RFH patients were initially assigned to one of two categories, suspected recent infection and established infection, based on a HIV antibody avidity index cut-off of 0.75.

Because late stage HIV can also result in a low avidity index (Chargelegue, Stanley et al. 1995), patients with avidity ≤0.75 were only retained for further analysis if they fulfilled one or a combination of the following criteria:

i. no evidence of AIDS defining illness at diagnosis, and
ii. last negative test <12 months prior to diagnosis, and/or
iii. documented seroconversion illness, and/or
iv. evidence of high risk exposure event(s) in the previous 6 months, and/or
v. confirmatory SOD result <1.0 obtained from 'detuned' assay

A minimum viral load of 1000 copies/ml for the sample used for *pol* sequence generation was imposed to ensure a sufficient number of viral particles were input into the PCR to enable observation of a quasi-species in Sanger sequence electropherograms. Briefly, 1ml of plasma was input into the extraction, nucleic acid was eluted into 25ul of elution buffer and 10ul of this was input into the reverse transcription polymerase chain reaction (RT-PCR) step. Assuming an arbitrary worst case scenario extraction efficiency of 65% and reverse transcription step efficiency of 10%, then 1000 copies/ml would become 650 copies of virus extracted into 25ul of elution buffer, and therefore 260 copies of virus input into the RT-PCR step. At 10% efficiency this would generate 26 cDNA molecules for PCR. Even allowing for a PCR efficiency of 50%, 13 molecules of cDNA should enable a 20-30% quasi-species to be detected in an electropherogram. Using the percentage of ambiguous nucleotides across *pol* for both patient categories, 10-fold cross-validation was used to select the cut-off that maximised the Youden index (a statistic which finds the cut-off value where a gain (or a loss) in specificity results in a loss (or a gain) of the same amplitude in sensitivity) (Youden 1950). Receiver Operating Characteristics (ROC) curve analysis was used to assess cut-off

performance. The process of cut-off selection was repeated with all non-B subtypes removed from the analysis to investigate subtype specific influences on classifier cut-off selection.

3.2.6 Validation of the nucleotide ambiguity cut-off in additional cohorts

In order to validate the performance of the nucleotide ambiguity cut-off, the UK Register of HIV Seroconverters and SMH cohorts were filtered to remove patients with uncertain infection stage categorisation, to produce a more stringent cohort for validation. The following criteria were used:

UK Register of HIV Seroconverters stringent cohort:

- Patients were considered to have had a recent infection at diagnosis if they were found to have laboratory evidence of seroconversion (HIV antibody negative with positive RT-PCR; test "incident" at low level (standard optical density < 1.0) using detuned assay (must be subtype B); equivocal HIV antibody test supported by a repeat test within a 2-week period showing a rising optical density; clinical manifestations of symptomatic HIV seroconversion illness supported by antigen positivity and <4 bands positive on Western Blot), or with a HIV positive antibody test within 125 days of an HIV negative antibody test.
- Patients were considered to have had an established infection at diagnosis if they had no contraindicative laboratory evidence and had a HIV positive antibody test greater than 125 days after a HIV negative antibody test.

SMH stringent cohort:

- Patients were considered to have had a recent infection at diagnosis if they had an avidity index of ≤0.75 or 'detuned' assay SOD result <1.0.

- Patients were considered to have had an established infection at diagnosis if they had an avidity index of >0.75 or indication of established infection through an alternative detuned assay.

After categorising patients into those diagnosed with recent infection and those diagnosed with established infection, a further filter of a minimum of 800bp sequence length was imposed in order to avoid classification based upon low sequence ambiguity generated by virtue of short sequence length (1 IUPAC coded base in a 800bp amplicon translates to 0.125% ambiguous nucleotides across the sequence, 2 translates to 0.25%). A count of nucleotide ambiguities was then made for each patient derived *pol* sequence, and the RFH cohort cut-off applied. A sensitivity and specificity analysis was then performed on each cohort to investigate cut-off performance. The sensitivity and specificity performance was repeated with all non-B subtype sequences removed.

3.2.7 Clinical data analysis

In order to further investigate the performance of the ambiguity cut-offs on the validation cohorts, for each nucleotide ambiguity cut-off and each cohort, a Student's t-test (2-tailed, type 3) was used to compare the $CD4^+$ cell counts and viral loads of the cohort patients categorised as recent infections using laboratory methodology (i.e. the original stringent classification outlined in the methods section) to the $CD4^+$ cell count and viral loads of the same patient cohort categorised as recent infections on the basis of the nucleotide ambiguity cut-off alone. This was repeated for those cohort patients categorised as established infections.

3.2.8 Comparison of Sanger sequence ambiguous nucleotides to deep-sequencing

A subset of the patients already identified as having been diagnosed with recent infection on the basis of avidity and clinical information had sections of their *pol* and *env* deep-sequenced using 454 pyrosequencing as previously described (section 2.10). Patients were selected on the basis of having plasma samples with ≥20000 HIV-1 RNA copies/ml available. Reads were quality filtered using the fastq filtering step within the UPARSE pipeline (Edgar 2013).

A comparison was made between the consensus sequence ambiguous nucleotides and deep-sequencing results to gain a greater understanding of the viral dynamics that result in mixed nucleotides. Briefly, a per-nucleotide 'consensus' sequence was constructed from the deep-sequencing reads alignment by selecting the most abundant nucleotide at each position. This consensus sequence was then aligned and compared to the Sanger *pol* sequence from the same sample and scored for a) exact nucleotide matches, b) if there was an International Union of Pure and Applied Chemistry (IUPAC) ambiguity code in the Sanger sequence at the position in question then the nucleotide variants above a prevalence cut-off of 2% in the deep-sequencing alignment were used to construct an IUPAC ambiguity code, and this was compared to the Sanger IUPAC code for a match, c) if the Sanger *pol* sequence had an A, C, G or T at the position in question, but this did not match the most abundant nucleotide in the deep-sequencing, then the second most abundant nucleotide present in the deep-sequencing alignment was compared to see if it matched, d) all other mismatches were scored as non-matches.

In order to ascertain whether or not patients were likely to have been infected by or to have developed more than one viral subpopulation over time, which would impact upon the applicability of the classifier, the deep-sequencing reads for each sample were passed through the UPARSE read clustering pipeline. Initial ‘seed’ subpopulation sequences were identified as subpopulation centroids on the basis of read abundance and being >3% genetically distant

from other centroids. Remaining reads with ≥97% identity to any subpopulation centroid were then clustered with that subpopulation. Sensitivity analysis was carried out to observe whether or not altering the distance cut-off would change the number of subpopulations identified within each sample. Two further cut-offs of 1% and 2% were substituted into the original pipeline, with all other parameters remaining unaltered. Read counts for subpopulations identified using the 3% cut-off were subjected to a margin of error correction using a 99% confidence interval (section 2.12). The lower boundary margin of error corrected percentage was used to decide whether the subpopulation was present at abundance greater than the chosen read abundance cut-off of 1%. The Poisson-Fitter tool was used to give an estimate of infection date using 500 randomly selected reads from the most abundant subpopulation in patient first time point samples only, to avoid over-estimates of infection date cause by multiple subpopulations. Where samples met the assumption of a star-like phylogeny, the estimate of time since most common recent ancestor (MRCA) was taken to be equivalent to the date of infection. The median time to MRCA across samples meeting the assumption of star-like phylogeny was used for the remaining samples where the assumption was not met.

3.3 Results

3.3.1 Classifier development

Using a guanidine-based avidity assay index cut-off of ≤0.75, 131/517 (25.3%) RFH patients were classified as having undergone seroconversion within the previous 125 days (95% CI: 85 to 164 days) and 386/517 (74.7%) patients were classified as having had an established infection at time of diagnosis. By case record review, 103/131 (78.6%) patients with avidity indices ≤0.75 had a history of seroconversion illness or high risk behaviour within the previous 6 months; 28/131 (21.4%) patients were excluded from the analysis because of a lack of such evidence, a $CD4^+$ cell count below 150 copies/ml, or they presented with an AIDS defining illness, which is associated with a reduced antibody avidity index (Chargelegue, Stanley et al. 1995) (Table 3.2). An additional 5 individuals were excluded from this patient group on the basis of low viral load (<1000 copies/ml), leaving 98 patients with a median viral load of 64900 copies/ml (IQR: 17268-311454 copies/ml). In addition, 17/386 patients with established infection at diagnosis were also excluded on the basis of low viral load, leaving 369 patients in this category. Patient *pol* sequences were stripped of codons associated with transmitted drug resistance mutations and the percentage of IUPAC nucleotide ambiguity coded nucleotides was correlated with avidity index using only the 98 patients with verified recent infection and the 369 patients with avidity indices >0.75.

RFH cohort, all subtypes: A cut-off of ≤0.17% ambiguous bases was found to maximise both sensitivity and specificity when used to identify sequences likely to have been generated from an individual who underwent seroconversion within the previous 125 days (95% CI: 85 to 164 days). The cut-off gave an area under the curve (AUC) of 0.83, and recent infection identification sensitivity 83/98 (85%) and specificity of 302/369 (82%). 83/150 individuals with a nucleotide ambiguity score

≤0.17% were clinically verified as recent infections. As the selection of patients was effectively cross-sectional, an estimate of prevalence of recent infection can be determined, allowing a positive predictive value of 0.85 to be calculated (Altman and Bland 1994).

RFH cohort, subtype B only: A cut-off based solely on subtype B sequences was identified in the same way, using 83 patients diagnosed with recent infection by the specified criteria, and 176 diagnosed with established infection. A cut-off of 0.14% was identified as maximising both sensitivity and specificity (63/83, 76% and 137/176, 78% respectively), with an AUC of 0.77, and positive predictive value of 0.85.

A more stringent cut-off of 0.00% nucleotide ambiguities was investigated to see if the performance of the classifier was significantly altered.

RFH cohort, all subtypes, 0.00% cut-off: On the full cohort the cut-off gave a sensitivity and specificity of 62/98 (63%) and 350/369 (95%) respectively, with a positive predictive value of 0.98.

RFH cohort, subtype B only, 0.00% cut-off: The subtype B only cohort gave a sensitivity and specificity of 48/83 (58%) and 161/176 (91%) respectively, with a positive predictive value of 0.96.

Exclusion criteria	Number excluded
No last negative test <12 months prior to diagnosis AND no documented seroconversion illness AND no evidence of high risk exposure event(s) in the previous 6 months AND no confirmatory SOD result <1.0 obtained from 'detuned' assay	13
No last negative test <12 months prior to diagnosis AND no documented seroconversion illness AND no confirmatory SOD result <1.0 obtained from 'detuned' assay	5
No last negative test <12 months prior to diagnosis AND no evidence of high risk exposure event(s) in the previous 6 months AND no confirmatory SOD result <1.0 obtained from 'detuned' assay	5
No last negative test <12 months prior to diagnosis AND no confirmatory SOD result <1.0 obtained from 'detuned' assay	1
No documented seroconversion illness AND no evidence of high risk exposure event(s) in the previous 6 months AND no confirmatory SOD result <1.0 obtained from 'detuned' assay	3
No evidence of high risk exposure event(s) in the previous 6 months AND no confirmatory SOD result <1.0 obtained from 'detuned' assay	1

Table 3.2. Table shows numbers of RFH nucleotide ambiguity development cohort patients with avidity indices ≤0.75 excluded on the basis of a combination of clinical criteria.

3.3.2 Validation

Validation of the cut-offs was carried out on two additional HIV-1 new diagnosis cohorts obtained from the UK Register of Seroconverters and St Mary's Hospital. Of the original 670 patients, 31 were removed from the UK Register of Seroconverters cohort due to sequence length falling below the 800bp length restriction, leaving 639 patients for the validation, 10 of the 169 original patients from the St Mary's cohort were removed for the same reason, leaving 159 patients for the validation.

UK Register of HIV Seroconverters and SMH cohorts, all subtypes, 0.17% cut-off: Applying the 0.17% cut-off to the UK Seroconverters Register cohort resulted in a sensitivity and specificity of 126/204 (62%), 258/435 (59%) respectively, and a positive predictive value of 0.52. For the SMH cohort the results for sensitivity and specificity were 17/27 (63%) and 85/132 (64%) respectively, with a positive predictive value of 0.39 (Table 3.3).

UK Register of HIV Seroconverters and SMH cohorts, subtype B only, 0.14% cut-off: The results when using the 0.14% cut-off on subtype B only sequences for both cohorts were a sensitivity and specificity of 101/170 (59%) and 259/395 (66%) respectively and positive predictive value of 0.56 for the UK Seroconverters Register, and for the SMH cohort, a sensitivity and specificity of 15/24 (63%) and 50/89 (56%) respectively and positive predictive value of 0.35.

UK Register of HIV Seroconverters and SMH full and subtype B only stringent cohort, 0.00% cut-off: After appropriate removal of all sequences below 800bp in

length, application of a 0.00% cut-off for the full UK Seroconverters Register cohort resulted in a sensitivity and specificity of 105/204 (51%) and 336/435 (77%) respectively and positive predictive value of 0.71, and for the subtype B only cohort 86/170 (51%) and 307/395 (78%) respectively and positive predictive value of 0.69. For the SMH full cohort the sensitivity and specificity was 10/27 (37%) and 106/132 (80%) respectively and positive predictive value of 0.42, and for the subtype B only cohort 9/24 (38%) and 69/89 (78%) respectively and positive predictive value of 0.43 (Table 3.4).

UK Serocon stringent 0.17%			
	Recent	**Established**	
≤0.17% ambiguities	126	177	303
>0.17% ambiguities	78	258	336
	204	435	
	0.62	0.59	

SMH stringent 0.17%			
	Recent	**Established**	
≤0.17% ambiguities	17	47	64
>0.17% ambiguities	10	85	95
	27	132	
	0.63	0.64	

Table 3.3. Sensitivity and specificity results for the 0.17% nucleotide ambiguity cut-off when applied to the UK Register of HIV Seroconverters and SMH cohorts. Columns 'Recent' and 'Established' correspond to categorisation using criteria outlined in methods section. Rows '≤0.17% ambiguities' and '>0.17% ambiguities' correspond to categorisation using the percentage of ambiguous nucleotides across the length of the *pol* sequence.

UK Serocon stringent

0.00%

	Recent	Established	
≤0.00% ambiguities	105	99	204
>0.00% ambiguities	99	336	435
	204	435	
	0.51	0.77	

SMH stringent

0.00%

	Recent	Established	
≤0.00% ambiguities	10	26	36
>0.00% ambiguities	17	106	123
	27	132	
	0.37	0.80	

UK Serocon stringent B only

0.00%

	Recent	Established	
0.00% ambiguities	86	88	174
>0.00% ambiguities	84	307	391
	170	395	
	0.51	0.78	

SMH stringent B only

0.00% cut-off

	Recent	Established	
0.00% ambiguities	9	20	29
>0.00% ambiguities	15	69	84
	24	89	
	0.38	0.78	

Table 3.4. Sensitivity and specificity results for the 0.00% nucleotide ambiguity cut-off when applied to the UK Register of HIV Seroconverters and SMH cohorts containing all subtypes (top panels), and containing patients with subtype B infections only (bottom panels).

3.3.3 Clinical data analysis

$CD4^+$ cell counts and viral loads of the cohort patients categorised as recent infections using laboratory methodology were compared to the $CD4^+$ cell count and viral loads of the same patient cohort categorised as recent infections on the basis of the nucleotide ambiguity cut-off alone. This was repeated for those cohort patients categorised as established infections. 313/670 (47%) of UK Register of HIV Seroconverter patients had $CD4^+$ cell count results obtained from blood samples taken > 1 month from the date of the blood sample the sequencing was carried out on, and these patients were excluded from the analysis. For all the RFH and SMH cohorts, $CD4^+$ cell count medians for patients whose HIV-1 infection lengths were categorised on the basis of the nucleotide ambiguity classifier did not differ significantly from the $CD4^+$ cell count medians of patients when categorisation of infection

length was based on laboratory evidence. For the UK Register of HIV Seroconverter cohort, a significant difference was found between the median $CD4^+$ count of those categorised as established infections on the basis of laboratory evidence and those categorised on the basis of the 0.17% cut-off. There was also a significant difference between the median viral loads of those categorised as recent infections on the basis of laboratory evidence and those categorised on the basis of the 0.17% cut-off. All other comparisons indicated no significant differences between median $CD4^+$ counts and median viral loads in the UK Register of HIV Seroconverter cohort (Table 3.5).

		Mean CD4 cell count (cells/mm3)					
		Recents			Establisheds		
Cohort	**Nucleotide ambiguity cut-off level**	**Laboratory classified**	**Ambiguity classified**	**p-value**	**Laboratory classified**	**Ambiguity classified**	**p-value**
Royal Free Hospital stringent, viral load filtered	0.17	562	518	0.13	254	227	0.14
Royal Free Hospital stringent, viral load filtered	0.00	562	560	0.71	254	270	0.47
UK Seroconverter stringent	0.17	440	477	0.09	**477**	**400**	**<0.05**
UK Seroconverter stringent	0.00	440	473	0.15	477	443	0.24
St Mary's Hospital stringent	0.17	665	520	0.05	400	340	0.44
St Mary's Hospital stringent	0.00	665	495	0.11	400	395	0.71

		Mean viral load (copies/ml)					
		Recents			Establisheds		
Cohort	**Nucleotide ambiguity cut-off level**	**Laboratory classified**	**Ambiguity classified**	**p-value**	**Laboratory classified**	**Ambiguity classified**	**p-value**
Royal Free Hospital stringent, viral load filtered	0.17	90157	43386	0.11	65612	80951	0.24
Royal Free Hospital stringent, viral load filtered	0.00	90157	55105	0.72	65612	70383	0.82
UK Seroconverter stringent	0.17	**190839**	**100000**	**<0.01**	56359	69219	0.09
UK Seroconverter stringent	0.00	190839	118000	0.08	56359	60650	0.40
St Mary's Hospital stringent	0.17	25069	18507	0.91	19027	19906	0.75
St Mary's Hospital stringent	0.00	25069	18507	1.00	19027	19891	0.95

Table 3.5. Comparison of median $CD4^+$ cell count and viral load between patients categorised as either recent infections or established infections by i) laboratory methods or ii) nucleotide ambiguity cut-off, using Wilcoxon rank sum test with continuity correction.

3.3.4 Comparison of Sanger sequence ambiguous nucleotides to deep-sequencing

Deep-sequencing results for each patient time point were compared to the Sanger population *pol* sequence generated from the sample, where available. 34 patients had Sanger *pol* sequence available for one or more time points (62 samples overall). Sequences were manually edited using the SeqScape v2.6 software (Applied Biosystems, Warrington, United Kingdom) prior to reference to deep-sequencing results. When exact nucleotide matches between the deep-sequencing and Sanger sequences were compared, a median of 99.7% (IQR: 98.7,100%) of nucleotides matched across the deep-sequencing amplicon (~390bp). When IUPAC and second most abundant nucleotide matches were taken into account, agreement went up to a median of 100% of nucleotides (IQR:100,100%) across the amplicons. Of interest, 5 samples had nucleotides at an abundance of only 5-15% in the deep-sequencing alignment were involved in an IUPAC ambiguity code match, a level much lower than classically accepted as likely to be revealed in Sanger sequence electropherograms.

The presence of minority subpopulations within patient samples was assessed using a read-clustering approach, with reads clustered into subpopulations with centroid sequences >3% genetically distant from one another. Subpopulations were verified by selecting only those present at a margin of error corrected cut-off of ≥1% total sample reads. Those meeting this criteria were then further scrutinised using BLAST and neighbor-joining trees, to assess the possibility that they were the result of contamination by multiplex identifier tag switching, or other sources of error (Carlsen, Aas et al. 2012). After margin of error correction, 45/65 (69.2%) and 43/65 (66.2%) samples had single *pol* and *env* subpopulations respectively, indicating 20/65 (30.8%) and 22/65 (33.8%) samples had >1 subpopulation in *pol* and *env* respectively. Where these minority subpopulations matched to within a <1% genetic distance of another subpopulation, unless the matching subpopulation was seen in additional time points from the same patient, it was excluded from the analysis on the basis

that it could be the result of MID tag switching. This filtering step reduced the number of samples likely to have >1 subpopulation to 13/65 (20.0%) and 18/65 (27.8%) for *pol* and *env* respectively.

Application of the Poisson-Fitter tool to 500 randomly selected reads from the dominant subpopulation of the first time point for each patient found that 22/36 patient first time points fitted the assumption of a star-like phylogeny. The median time since MRCA for these samples was 120 days (IQR: 106.5-137 days), and this value was used for the patients not fitting the assumptions of star-like phylogeny. Pertinent to the potential utility of the nucleotide ambiguity tool, 4/36 (11.1%) patients had good evidence for multiple subpopulations in *pol* at first time point, all were assigned a time since MRCA of 120 days (0/4 patients met the assumption of star-like phylogeny required by Poisson-Fitter). For *env*, 8/36 (22.2%) patients had good evidence for multiple subpopulations, three of which matched samples with multiple subpopulations in *pol* (0/8 patients met the assumption of star-like phylogeny required by Poisson-Fitter).

3.4 Discussion

This study aimed to investigate the use of a HIV-1 *pol* sequence nucleotide ambiguity classifier to determine the length of HIV-1 infection based solely on sequence information obtained as part of routine drug resistance surveillance. The classifier was developed on a cohort of patients who had guanidine-based avidity assay results and clinical information available, enabling stringent categorisation of patients into recent and established infection categories prior to sequence ambiguity analysis. The classifier cut-off selected was ≤0.17% nucleotide ambiguities across the *pol* region, which was taken to be indicative of recent infection, where recent infection was defined, through linkage with the avidity assay, as seroconversion within the previous 125 days (95% confidence interval 85 to 164 days). This cut-off was selected to maximise both sensitivity and specificity, which were calculated as 85% and 82% respectively, with a positive predictive value of 0.85.

When validated against two independent cohorts of newly diagnosed patients (from the UK Register of HIV Seroconverters and St Mary's Hospital, London), the cut-off offered a reduced level of discrimination between recent and established infections, with sensitivity and specificity calculated as between 62-63% and 59-64% respectively, and a positive predictive value of 0.52 and 0.39 respectively. When the nucleotide ambiguity cut-off was reduced to 0.00% ambiguities across *pol* (with a minimum sequence length of 800bp), the positive predictive values were improved for both validation cohorts (0.71 and 0.42 respectively), but were still low when compared to the development cohort positive predictive value of 0.98 at this cut-off level.

Restricting the development cohort to the predominating subtype present, subtype B, and validating on only subtype B sequences from validation cohorts, did not alter the sensitivity and specificity to any great degree, regardless of whether the cut-off was developed to maximise sensitivity and specificity, or to maximise specificity only. This may

be a reflection of the predominance of subtype B in the development and validation cohorts, which varied from 71-88% subtype B. If this is the case, then it may be advisable to investigate potential subtype effects further in settings where subtype B is not the most common subtype present, as in these settings there may be a variety of factors that affect the immune response or the maturation curve dynamics of the underlying antigen avidity assay, such as HIV-1 subtype, genetic background and so forth (Sakarovitch, Rouet et al. 2007).

It is important to note that sample size of a cohort will inevitably have an effect on the confidence that the selection is representative of the true underlying proportion of recent infections in the population as a whole. The smaller the sample size the greater the margin of error in the number of recent infections selected will be. This can have impressive effects upon the positive prediction value. For example, a margin of error at a 99% confidence level for the 159 patients in the St Mary's cohort would mean that the underlying number of 'true' recent infections sampled could vary between 12 and 38 (the actual number sampled being 27). This in turn can lead to variation in the positive predictive value of 0.24-0.69 when using the 0.17% nucleotide ambiguity cut-off.

It is encouraging to observe that the clinical profile of RFH and SMH cohort patients identified as having recent infection using the nucleotide ambiguity classifier was not found to differ significantly from the patients with recent infection identified using laboratory methods, in terms of viral loads and $CD4^+$ cell count. However, there were found to be differences between the median $CD4^+$ count and median viral load for the UK Register of HIV Seroconverters cohort when using the 0.17% cut-off, which may indicate support for the use of the more stringent 0.00% cut-off.

The analysis of deep-sequencing results enabled exploration of the underlying quasi-species landscape that contributes to the consensus sequence seen when using traditional Sanger sequencing. It also enabled an assessment of the extent to which recent infections

consist of a single subpopulation that is generated by expansion of a single founder virion after infection – an assumption which, if violated in large numbers of patients, may impact upon the suitability of the nucleotide ambiguity methodology to determine length of infection. The high degree of concordance between deep-sequencing variants and Sanger sequencing nucleotide ambiguities suggests that Sanger sequencing is a good representation of the underlying quasi-species present within a patient, and that nucleotide ambiguities seen in electropherograms do represent the true underlying proportions of variants present when the minority species is at a certain minimum level of abundance. This validates the use of Sanger consensus sequences as a proxy for underlying genetic complexity in HIV-1 infections, an assumption which the use of nucleotide ambiguity implicitly makes.

It is important to note that the route of HIV-1 infection can confound nucleotide ambiguity based classifier accuracy due to the fact that infection routes associated with risk groups such as men who have sex with men (MSMs) and people who inject drugs (PWIDs) may enable a greater number of viral particles to found an infection, generating an immediate quasi-species (Bar, Li et al. 2010; Li, Bar et al. 2010). Co-infection, where an individual is infected by two or more individuals before the initial infection has had a chance to establish itself, may also lead to confounding nucleotide ambiguities in the viral sequence, but is thought to be comparatively rare in the overall HIV-1 infected population (Cornelissen, Jurriaans et al. 2007). The results of the read clustering approach taken in this study strongly indicate that the majority of infections in this cohort were founded by a single virion, or several highly genetically similar virions, as has been observed in studies in other populations (Zhang, MacKenzie et al. 1993; Zhu, Mo et al. 1993; Keele, Giorgi et al. 2008). Of the 9 individuals with >1 subpopulation in *pol* or *env* in their first time point, 3 individuals had support for 2 subpopulations in both *pol* and *env*, 5 had evidence of 2 subpopulations in *env* but not in *pol*, and 1 had support for 2 subpopulations in *pol* but only 1 in *env*. This perhaps

suggests that *env* may be under greater pressure to diversify under host immune pressure, such that even if one virion founded the infection, immune pressure is rapidly brought to bear, and evolution to evade this pressure occurs in *env* first. Overall, the low number of patients in this cohort with evidence for >1 subpopulation in the time point closest to infection lends legitimacy to the application of the nucleotide ambiguity classifier to patients newly diagnosed with HIV-1. However, it may well be worth more thoroughly investigating patients likely to have been infected by different transmission routes (e.g. MSMs and PWIDs), and who may therefore be subject to differing infecting inoculums, as it may not be appropriate to apply the ambiguity classifier to their viral sequences.

The three other similar studies on this topic produced nucleotide ambiguity cut-offs of >0.50% for infections >1 year (Kouyos, von Wyl et al. 2011), <0.45% for infections <155 days (Ragonnet-Cronin, Aris-Brosou et al. 2012), and <0.47% for infections <1 year (Andersson, Shao et al. 2013). It is worth noting that across 1300bp (the typical length of *pol* sequenced for drug resistance surveillance), the overall range, including the results of this chapter, of 0.17% to 0.50% translates to between 2 and 7 ambiguous nucleotides. Whilst the variability of the cut-offs and the definitions of the recent infection period do differ, the combined message from these previous studies and the results of this chapter seems to be that nucleotide ambiguity does increase over the initial period of infection, and as such it can be used to identify individuals with recent infection with a sensitivity and specificity approaching that of laboratory tests.

In terms of issues and limitations of the study, it is important to acknowledge that differences in sequencing and sequence analysis performance across different clinical laboratories have the potential to impact upon the applicability of the classifier across sequences generated by different laboratories. The potential impact of this issue could be further investigated by performing repeats of sequencing runs to check for the consistency of

ambiguous peaks between experiments. The extent to which the same individual calls ambiguous nucleotides at different time points, and how different individuals call the same ambiguous nucleotides could also be ascertained experimentally. Reassurance that the potential significance of this factor may not have an overly negative impact upon the suitability of the nucleotide ambiguity approach comes from an investigation into variation in DNA sequencing performance across laboratories carried out by Patton, Wallace et al. (Patton, Wallace et al. 2006). The study used a set of reference DNA samples (fragments of exon 10 of the cystic fibrosis transmembrane conductance regulator (CFTR) gene) to investigate laboratory differences in the quality of sequence data, ability to detect variant genotypes, and mutation nomenclature. They found a 5% genotyping error rate, split roughly equal between false negative and false positive errors. In their judgement, the majority of the sixty-one laboratories included in the analysis "produced results of acceptable diagnostic quality as judged by these indicators" and suggested that where errors did occur, it was primarily down to "human error, failure of mutation scanning software, or a combination of both".

With this last comment in mind, and with the caveat that Patton, Wallace et al. (2006) did not aim to compare performance in identification of ambiguous nucleotides in sequencing, it is arguable that a nucleotide ambiguity cut-off of 0.00% may reduce the significance of sequence quality issues in this context (in combination with a minimum sequence length requirement). Sequence quality issues, though having the potential to increase false positive ambiguous peaks (whether human or software read), should have less of an effect on specificity: intuitively, it seems unlikely that poor quality sequence above a certain length would have no ambiguous nucleotides, and hence unlikely that a sequence with more than 0.00% ambiguous nucleotides – poor quality or not – would be classified recent by mistake. The minimum sequence length of 800bp imposed in this study means that, on

average, the presence of zero ambiguous nucleotides should imply an ambiguous nucleotide rate of no more than 0.0625% (because 0.5 ambiguous nucleotides across 800bp is 0.0625%, and, probabilistically, at this rate one in every two sequences should harbour an ambiguous nucleotide). Substantially shorter sequences with zero ambiguous nucleotides could lead to an incorrect estimation of nucleotide ambiguity – e.g. if there are zero ambiguous nucleotides in 400bp of truncated *pol* sequence, it is impossible to know if the full 800bp *pol* sequence would have had zero or ≥1 nucleotide ambiguities. The 800bp minimum sequence length also allows laboratories utilising the TRUGENE HIV-1 Genotyping Assay to apply the nucleotide ambiguity classifier to their archived sequences, as the assay generates sequence for codons 1-99 of protease and codons 41-237 of reverse transcriptase (i.e. 885bp of *pol*) (Grant, Kuritzkes et al. 2003). Further work on the impact laboratory and protocol variation may have upon sequence quality may be warranted.

A final point must be made in relation to the use of the guanidine-based avidity assay as the reference standard, whilst the assay itself is not a gold-standard. This is, of course, not without precedent in the development of clinical markers, where there may often not be a single gold-standard reference test available, and a variety of methods have been proposed to address this (Spiegelman, Schneeweiss et al. 1997; Hawkins, Garrett et al. 2001; Rutjes, Reitsma et al. 2007). It may be worth investigating patients where the guanidine-based avidity assay and nucleotide ambiguity results are discordant, and employ some other indicators of infection length such as $CD4^+$ cell count, in order to ascertain the likely cause of the discrepancy. In terms of an accurate assessment of overall sensitivity, specificity and positive predictive value that take the sensitivity and specificity of the guanidine-based assay into account, various methods for carrying this out in a statistically robust manner remain under development, and so it is beyond the scope of this study to fully resolve this limitation. Whilst this may impact upon the use of the nucleotide ambiguity classifier in isolation in a

clinical setting, the classifier arguably remains useful in the context of additional clinical data in individual patient scenarios, and is likely to be a useful research tool when applied to large-scale patient cohorts, with limited clinical data available to complement their viral sequence data.

In conclusion, we have investigated population sequence nucleotide ambiguity as a measure of the length of HIV-1 infection by linking virus sequence with results from a guanidine-based avidity assay able to identify individuals who have seroconverted within the previous 125 days. As noted by previous investigators in this area, the classifier has the potential to be useful in a clinical setting as part of an overall recent infection testing algorithm, when applied in conjunction with additional data, such as $CD4^+$ cell count. The classifier may also be useful on its own when utilising sequence data obtained from national drug resistance surveillance databases, where additional clinical data may not be available for the majority of patients. In this context, the classifier has the potential to be combined with phylogenetic approaches to investigate the role of the initial stage of HIV-1 infection in larger-scale epidemic transmission dynamics.

Chapter Four

An investigation into the extent to which individuals with recent infection drive the formation of HIV-1 transmission clusters in the UK

4 An investigation into the extent to which individuals with recent infection drive the formation of HIV-1 transmission clusters in the UK

Abstract

A significant proportion of HIV-1 transmission within the UK epidemic may come from individuals in the recent phase of infection due to particular risk factors associated with this risk group. Phylogenetic approaches can shed light on the extent to which this is true, and this chapter represents a large scale phylogenetic analysis of infected individuals in the UK, in combination with application of a nucleotide ambiguity based marker of infection length developed in chapter three.

Within the UK HIV Drug Resistance Database, 5250 transmission clusters were identified using a bootstrap support cut-off of ≥0.95 and a maximum intra-cluster pairwise genetic distance ≤0.045 substitutions per nucleotide (Hamming distance). A disproportionate number of individuals classified as having recent infection at blood sample collection date were found to be linked to transmission clusters. Using two proof-of-principle modelling approaches, there was evidence that larger clusters had a greater proportion of transmissions from the recent phase of infection.

The approach taken within the chapter lends further support to the hypothesis that individuals with recent infection play a disproportionate role in the UK HIV-1 epidemic. The work also demonstrated the utility of a consensus sequence based marker of infection length for the investigation of epidemic dynamics.

4.1 Introduction

The initial stages of HIV-1 infection, around the period of seroconversion, are associated with extremely high levels of viral replication (Pilcher, Shugars et al. 2001; Fiebig, Wright et al. 2003), with individuals in this period frequently having plasma HIV-1 RNA levels in excess of 10^6 copies/ml. High virus levels in blood are mirrored in the elevated levels found in genital secretions (Pilcher, Joaki et al. 2007; Morrison, Demers et al. 2010), with a number of studies indicating that higher viral levels in the genital compartment increase the probability of transmission to sexual partners per coital act (Leynaert, Downs et al. 1998; Hollingsworth, Anderson et al. 2008).

Mathematical models incorporating this higher level of infectivity with particular patterns of sexual behaviour during the early stage of infection, when individuals may not be aware of their infection status, suggest that individuals with recent infection play a disproportionate role in the spread of national HIV-1 epidemics (Jacquez, Koopman et al. 1994; Koopman, Jacquez et al. 1997). These studies, however, have lacked important data on contact networks, such as the rate at which individuals form sexual relationships with others.

More recently, there have been attempts to use a phylogenetic approach to gain a greater understanding of transmission events and contact networks. These are captured as statistically robust clusters of individuals with highly related viral strains, representing chains of infections more closely linked to each other than the average level of linkage across the epidemic as a whole. Once identified, factors leading to the formation of these more highly related clusters, such as the stage of HIV-1 infection, can be assessed (Lewis, Hughes et al. 2008; Fisher, Pao et al. 2010). To date, these studies have supported the idea of a disproportionate role in transmission for individuals with recent infection. The studies have addressed particular HIV epidemics of limited size, where transmission intervals were assessed using either serological data available in individual clinics, or methodologies that

would be computationally unfeasible with the large numbers of sequences involved on the scale of whole epidemics (Lewis, Hughes et al. 2008; Fisher, Pao et al. 2010).

This study represents an extension of the phylogenetic approach, by investigating a large-scale national HIV-1 phylogeny using viral samples obtained from HIV-1 infected individuals from across the UK between 1997 and 2009. The analysis incorporates a previously developed classifier for use with nucleotide sequence data which categorises individuals into two categories: individuals with recent infection at blood sample collection date and individuals with established infection at blood sample collection date. The classification is based upon the proportion of ambiguous nucleotides present in the viral sequence of the individual, and the period considered to be recent is 125 days post-seroconversion (95% CI: 85 to 164 days). Transmission clusters were identified using standard phylogenetic methodology, and individuals within those clusters categorised into those having had a recent infection at time of blood sample collection, and those not, and their contribution to transmission within the cluster assessed. The analysis was then extended to include transmitted drug resistance circulating within this group, as this may also have important consequences for healthcare intervention strategies.

Two proof-of-concept methods were investigated to demonstrate the potential utility of applying the nucleotide ambiguity classifier to sequences within a phylogeny in order to shed light on the transmission dynamics of individuals with recent infection within clusters. Both methods utilised a combination of information captured by phylogenetic analysis and results of the application of the nucleotide ambiguity classifier to examine the extent to which individuals in the initial stages of HIV infection may disproportionately contribute to onward transmission of HIV in the UK epidemic.

4.2 Materials and Methods

4.2.1 Identifying transmission clusters within UK HIV Drug Resistance Database

The UK HIV Drug Resistance Database (UKHIVDRD), established in 2001, is a central repository for drug resistance tests performed as part of routine HIV clinical care throughout the UK. The UKHIVDRD release used in this study contained 55556 *pol* gene sequences collected from 43002 anonymised individuals.

Sequences were supplied individually pre-aligned to a reference sequence (consensus B) using the LAP alignment program (Huang and Zhang 1996). Where duplicate sequences from the same patient existed, the chronologically earliest sequence was used. For practicality of analysis, sequences were divided into subtype subsets based on the REGA subtyping algorithm (De Oliveira, Deforche et al. 2005). To avoid possible convergent evolution due to transmitted drug resistance associated mutations (DRAMs) at particular codons, phylogenies were constructed using FastTree2.1 (Price, Dehal et al. 2010) with i) alignments with DRAM associated codons removed, and ii) alignments with only the 1st and 2nd nucleotide positions from each codon removed. Bootstrapping was carried out using 1000 replicates with FastTree2.1.

Clustering of particular sequences was deemed to be robust if there was bootstrap support of ≥0.95 in both DRAM codon stripped and 3rd codon position only trees, and a maximum intra-cluster pairwise genetic distance ≤0.045 substitutions per nucleotide (Hamming distance). Nested clusters were removed, so that all sequences linked to clusters were linked to one cluster only. As transmission clusters comprising individuals from the UK were the primary focus of the study, clusters from each subtype subset were aligned with relevant subtype sequences from the Los Alamos database (with UK sampled sequences removed) and trees re-drawn in order to identify and remove or split clusters that were linked to non-UK sequences with Shimodaira-Hasegawa-like local support values of ≥0.99

(Shimodaira and Hasegawa 1999). All sequences were assessed for treatment status using information submitted to the UKHIVDRD from clinics, and for evidence of drug resistance using the Sierra web tool (http://hivdb.stanford.edu/pages/webservices/).

4.2.2 Assessing contribution of individuals with recent infection and established infection to transmission clusters

In order to assess the role of individuals categorised as having been recently infected at time of blood sample collection in the formation of transmission clusters, two levels of focus were taken; one with a reduced set of individuals that had greater laboratory and clinical characterisation, allowing a more stringent categorization of either recent or established infection (*New diagnosis cohort)*; the second with a UK-wide population that had more limited laboratory and clinical data available (*Overall UK HIV Drug Resistance Database).*

For the *New diagnosis cohort*, laboratory and clinical data and viral sequences stripped of personal identifiable information were obtained from 1570 individuals newly diagnosed with HIV-1, whose data were held either at The Royal Free and St Mary's Hospitals in London, or the UK Register of HIV Seroconverters: a UK-wide database of individuals with documented evidence of seroconversion. Individuals were categorized as having a recent or established infection at the time of HIV diagnosis on the basis of either

i. laboratory based serology,
ii. a combination of last negative/first positive HIV test date together with application of a nucleotide ambiguity classifier based upon viral sequence nucleotide ambiguity, or
iii. application of the nucleotide ambiguity classifier alone, where no other data was available

New diagnosis cohort viral sequences already submitted to the UKHIVDRD as part of routine national HIV anti-retroviral drug resistance surveillance procedures were identified to avoid duplication.

Secondly, a nucleotide ambiguity classifier based upon viral sequence nucleotide ambiguity alone was applied to all treatment naive individuals in the UKHIVDRD, in an attempt to gain a larger picture of the contribution to onward transmission from each infection category across different subtypes and cluster sizes.

Previous work developing the nucleotide ambiguity classifier investigated two cut-offs for categorisation of infection length, a 0.17% nucleotide ambiguity cut-off that aimed to maximise both the sensitivity and specificity of categorisation of individuals with recent infection at blood sample collection date, and a second, more stringent cut-off of 0.00% nucleotide ambiguities across ≥800bp sequence, which was previously found to increase the positive predictive value of the classifier – both cut-offs were applied within this study (chapter three).

4.2.3 Modelling recent infection dynamics within clusters

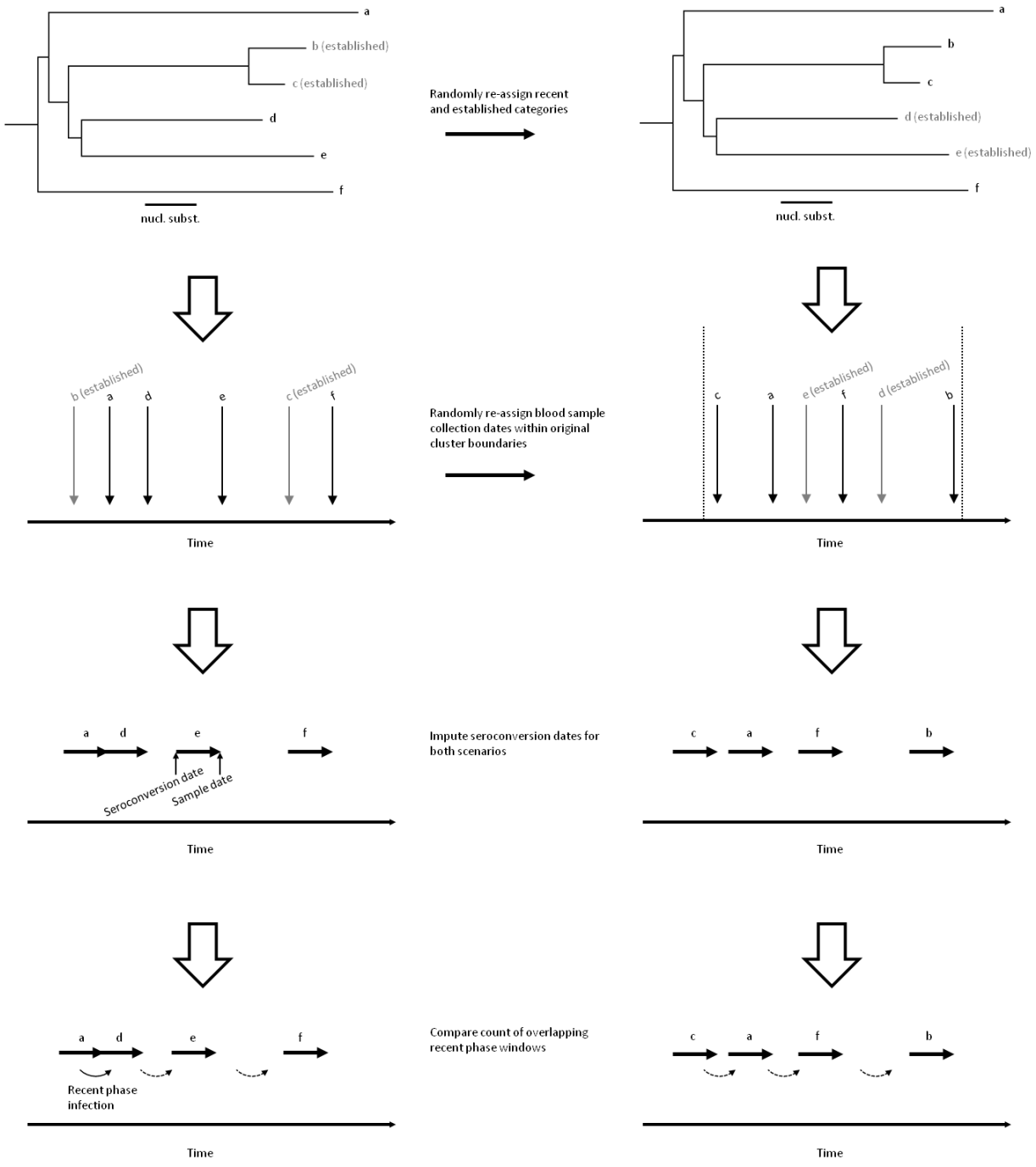
A model of infection transmission within clusters was developed in an attempt to describe the extent to which individuals categorised as having had recent infection at blood sample collection date contribute to onward transmission close to their infection. The model was set-up as follows: firstly, it was necessary to discard individuals categorised as having had established infections at time of blood sample collection, where according to the assumptions of the nucleotide ambiguity classifier, it is not possible to put a minimum-bound on the date of infection (i.e. established infections are only known to have *not* seroconverted in the previous 125 days). Next, a date of seroconversion was inferred for each individual with recent infection by subtracting 125 days from their blood sample collection date, thereby

allowing the recent infection to be ordered with respect to calendar date of infection. This in turn enabled analysis of the extent to which those still close to their date of infection (which is likely to have occurred 2-3 weeks prior to seroconversion) may drive transmission within an epidemic: briefly, a count of the number of infections coming from within the 125 day recent infection window was carried out by assuming each infection was generated by the infection immediately upstream of it, i.e. a consecutive date ordered approach (Fig 4.1.). This simplification assumes each individual may transmit to no more than one individual in the cluster (apart from the chronologically last individual, who is the last in the chain), and likely serves to minimise the time between transmissions, potentially leading to an over-estimation of the number of transmissions from the recent phase of infection. To address this issue, the model compares this count of the number of infections coming from within the 125 day recent infection window for any particular cluster to an equivalent count based on a cluster reconstructed to match the original in terms of size and proportion of recent infections, but with randomly assigned seroconversion dates. It is postulated that if such a comparison reveals that UK phylogeny based transmission clusters display a greater degree of overlap of recent phase infections compared to equivalent clusters with randomly assigned dates of seroconversion (i.e. not based on actual blood sample dates), it may imply that transmission clusters across the UK have a higher rate of individuals in the recent phase of their HIV-1 infections linked to others also in the recent phase of their infection than may be expected by chance. This in turn would support the paradigm that transmission from the recent phase of infection is the primary driver in the formation of such clusters.

In more detail, for each cluster identified within the UK phylogeny, an equivalent cluster was reconstructed, whereby the infections within the original cluster were randomly re-assigned a seroconversion date somewhere between the oldest and youngest original cluster blood sample collection dates. A subset of these infections were then randomly

categorised as recent infections in a proportion matching the original cluster, and the degree of overlap between recent phase infections compared to that of the original cluster; in effect the model attempts to detect greater temporal grouping of recent infections (i.e. overlap of recent infection windows) than expected if having a recent HIV-1 infection makes an individual no more likely to transmit to another individual than having an established infection.

Figure 4.1. Schematic showing infection modelling approach. Top and second row: original cluster infections are mapped onto a timeline using sample dates from UKHIVDRD, or at random for the 'stochastic' scenario. Third row: seroconversion date is inferred from sample date for all recent infections; established infections are ignored because no infection date can be inferred. Bottom row: the consecutive infection scenario assumes each individual infects the individual immediately downstream.


Original cluster ordered infection scenario
Randomised date ordered infection scenario
a
b (established)
c (established)
d
e
f
nucl. subst.
Randomly re-assign recent and established categories
b (established)
a
d
e
c (established)
f
Time
Randomly re-assign blood sample collection dates within original cluster boundaries
c
a
e (established)
f
d (established)
b
Time
a
d
e
f
Seroconversion date
Sample date
Time
Impute seroconversion dates for both scenarios
c
a
f
b
Time
a
d
e
f
Recent phase infection
Time
Compare count of overlapping recent phase windows
c
a
f
b
Time

4.2.4 Counting recent phase transmissions using nucleotide ambiguity and pairwise genetic distance

An additional attempt to investigate the extent to which transmission clusters are driven by recent infection was made by combining the nucleotide ambiguity cut-off with pairwise nucleotide distance measures. A reasonable estimate of the rate of nucleotide evolution for HIV-1 was obtained from previous work (Leitner and Albert 1999), as $2.7 \pm 0.5 \times 10^{-3}$ in p17 and $6.7 \pm 2.1 \times 10^{-3}$ substitutions/site per year in the V3 loop. Although the V3 region is likely to be under greater selective pressure than *pol* in the initial stages of infection, the latter rate of evolution was taken to be a useful upper limit of genetic distance for the model. This rate translates into ~3.0 substitutions per 125 days across the 1300bp region of *pol* typically submitted to the UKHIVDRD and if the pair of sequences were also categorised as recent infections, on the basis of the nucleotide ambiguity classifier, and their 125 day recent infection windows overlapped, the putative transmission was counted as being from the recent phase of infection. Where multiple individuals overlapped in a single group in this manner, counts of recent phase infections were adjusted appropriately (i.e. a putative recent phase transmission event between patient A and patient B, and between B and patient C means that there cannot have been a putative recent phase transmission event between patient A and patient C).

It should be stressed that, as with the cluster reconstruction approach outlined above, this method does not attempt to identify actual transmission events between individuals with highly related sequences. In such pairings there remains the possibility that a third person, not captured within the UKHIVDRD, may have transmitted HIV to both individuals in question, and that these two individuals were diagnosed before the virus had significantly diverged from the original infecting virus. A further caveat is that the model cannot capture individuals who may have been part of a recent phase transmission event at some point in the past, but

were categorised as having been in the established phase of infection when sampled, or where the blood from two individuals involved in a recent phase transmission event was sampled at sufficiently distant time points that the virus had time to diverge beyond the pairwise distance threshold.

In light of these caveats, the method primarily attempts to enable comparisons between subtypes and between cluster sizes, in terms of the difference in the number of transmissions from individuals with recent infection. A greater or lesser proportion of transmission events identified as coming from the recent infection phases when comparing, for example, different categories of cluster size, may indicate the extent to which recent infection is playing a role within different parts of the UK HIV-1 epidemic.

4.3 Results

4.3.1 Study population characteristics

The initial sequence dataset was comprised of 55556 anonymised HIV *pol* sequences obtained from 43002 individuals in the United Kingdom between 1997 and 2010 Where multiple sequences existed for the same patient, the chronologically earliest sample was used for analysis, resulting in the removal of 12554 sequences. A subset 16362 of individuals within the UKHIVDRD had demographic data available as part of the UK Collaborative HIV Cohort (UK CHIC, www.ukchic.org.uk) (Table 4.1.).

The *pol* sequences covered the protease gene (297bp) and part of the reverse transcriptase gene (1023bp), and were subtyped using the REGA subtyping algorithm (De Oliveira, Deforche et al. 2005). A wide variety of subtypes and circulating recombinant forms were represented, with the main subtypes being present in the following proportions: A (5.8%), B (54.0%), C (23.4%), CRF01_AE (1.5%), CRF02_AG (3.9%), D (1.3%), G (1.9%), unassigned (6.9%) and the remaining 1.3% being made up of less common subtypes and recombinants forms.

	All UK CHIC patients
Female	3509 (21.5%)
Male	12848 (78.5%)
Blood products	70 (0.4%)
Heterosexual	4416 (27.7%)
Homo/bisexual	9179 (57.7%)
IDU	444 (2.8%)
Not known	1226 (7.7%)
Other	584 (3.7%)
Black-African	3173 (20.5%)
Black-Caribbean	443 (2.9%)
Black-other/unspecified	287 (1.9%)
Indian/Pakistani/Bangladeshi	155 (1%)
Not known	679 (4.4%)
Other	383 (2.5%)
Other Asian/Oriental	231 (1.5%)
Other/Mixed	571 (3.7%)
White	9541 (61.7%)

Table 4.1. Demographic breakdown of 16362 UK Collaborative HIV Cohort patients linked to the UK HIV Drug Resistance Database.

4.3.2 Identification of putative transmission clusters

All 55556 *pol* gene sequences collected from across the UK as part of routine transmitted drug resistance surveillance were aligned individually to a reference sequence (consensus B) using the LAP alignment program (Huang and Zhang 1996). 12554 sequences were removed from the alignment, and the remaining sequences were then divided into subtype subsets based on the REGA subtyping algorithm for computational practicality (with 2 individuals being excluded from the subsets on the basis of one patient having group O HIV-1 and one patient having HIV-2 according to REGA subtyping). Phylogenetic trees were constructed

using the FastTree2.1 phylogenetic algorithm utilising i) alignments stripped of codons linked to drug resistance associated mutations (DRAMs), and ii) alignments with the 1st and 2nd nucleotide positions of each codon removed. Bootstrapping was carried out in FastTree2.1 with 1000 replicates. Clusters of sequences with bootstrap support ≥95% and a maximum intra-cluster genetic distance ≤0.045 substitutions per nucleotide were deemed likely to represent groups of individuals with viruses significantly more closely related to each other than on average within the overall UKHIVDRD, in turn potentially identifying individuals involved in onward transmission of HIV due to particular underlying epidemiological or biological factors. In terms of the effect of nucleotide ambiguities on the identification of transmission clusters, FastTree2.1 treats ambiguous nucleotides as missing data and hence will ignore such sites. The Analyses of Phylogenetics and Evolution (APE) module in R (Paradis, Claude et al. 2004) was used to calculate the pairwise Hamming distances between sequences in clusters. The Hamming distance is simply the number of differences between two strings, and hence an ambiguous nucleotide at a particular site in one sequence compared to a pure nucleotide in the other sequence will be counted as a difference between the two sequences even if the ambiguous nucleotide incorporates the pure nucleotide (e.g. R vs. A). This was deemed to be unlikely to present a major issue for the methodology when the overall distribution of nucleotide ambiguities in the UKHIVDRD was considered: across the 43002 unique patient sequences, the median percentage ambiguity was 0.69% (90 CI: 0.00% to 2.06%), whereas a cluster would have to contain a sequence with a percentage ambiguity of at least 4.5% to exclude it from being identified as a potential transmission chain. Overall, the cluster definition criteria resulted in identification of 5250 potential transmission clusters, involving 16923/43002 (39.4%) individuals from the UKHIVDRD (Table 4.2.). Of these clusters, 1887 (36.0%) were made up of greater than two individuals (i.e. excluding transmission pairs).

Subtype	No. clusters	No. cluster > 2	No. clusters ≥ 10
A	305	78	2
B	2466	1066	130
C	1488	429	8
CRF01_AE	82	21	0
CRF02_AG	247	74	0
D	57	17	1
G	112	33	1
Non-major	86	28	2
U	407	141	9
Total	5250	1887	153

Table 4.2. Numbers of transmission clusters identified using ≥95% bootstrap and maximum intra-cluster pairwise genetic distance ≤0.045 substitutions per nucleotide, broken down by subtype and cluster size.

4.3.3 Analysis of transmission cluster proportions and clinical data

1. <u>New diagnosis cohort</u>

687/1570 (43.8%) individuals had the stage of their infection at diagnosis categorized on the basis of serology, 604/1570 (38.5%) by a combination of known last negative HIV test date first positive HIV test date and application of the nucleotide ambiguity classifier, and 279/1570 (17.8%) by application of the nucleotide ambiguity classifier alone. 1231/1570 (78.4%) of sequences from the new diagnosis cohort were part of the UKHIVDRD. Of these 1231 sequences, 519 (42.2%) belonged to a putative transmission cluster.

0.17% nucleotide ambiguity cut-off:

Overall, 381/1570 (24.3%) individuals were categorized as having had recent infection at diagnosis. 1291 individuals were classified on the basis of laboratory evidence of

seroconversion, or a combination of known last negative HIV test date first positive HIV test date and application of the 0.17% nucleotide ambiguity classifier, 285/1291 (22.1%) individuals were categorised as recent infections.

0.00% nucleotide ambiguity cut-off:

When using the 0.00% nucleotide ambiguity cut-off, 309/1539 (20.1%) individuals were categorized as having had recent infection at diagnosis. 1254 individuals were classified on the basis of laboratory evidence of seroconversion, or a combination of known last negative HIV test date first positive HIV test date and application of the 0.00% nucleotide ambiguity classifier, 253/1254 (20.2%) individuals were categorised as recent infections.

2. Overall UK HIV Drug Resistance Database

UKHIVDRD individuals classified as treatment experienced at their earliest sample date were deemed likely to be present in the database because they were experiencing virological failure, and it was assumed to be improbable that they had a recent infection at that time point. Accordingly, the nucleotide ambiguity classifier was only applied to UKHIVDRD individuals classified as treatment naive. Of the 43002 individuals within the UKHIVDRD, 23320 (54.2%) were classified as treatment naive at blood sample collection date (excluding one patient infection subtyped by REGA as HIV-2). Application of the 0.17% nucleotide ambiguity cut-off to the 23320 treatment naive individuals categorised 5975/23320 (25.6%) individuals as having had a recent infection at blood sample collection date.

0.17% nucleotide ambiguity cut-off:

For all subtype categories combined, treatment naive individuals classified as having had recent infection at blood sample collection date using the 0.17% cut-off were more likely to:

i) have higher median first and last (before treatment) $CD4^+$ cell counts (469 cells/mm^3 vs. 350 cells/mm^3, Mann-Whitney p-value < 0.0001 and 357 cells/mm^3 vs. 254 cells/mm^3, Mann-Whitney p-value < 0.0001 respectively),

ii) have lower median last viral load (before treatment) (27762 copies/ml vs. 41864, Mann-Whitney p-value < 0.0001),

iii) be male (Fisher's exact test p < 0.0001, OR 2.58 [CI 95: 2.27, 2.94]),

iv) be identified with the men who have sex with men (MSM) risk group (Fisher's exact test p < 0.0001, OR 1.82 [CI 95: 1.67, 2.00]), and

v) iv) have white ethnicity (Fisher's exact test p < 0.0001, OR 1.99 [CI 95: 1.81, 2.19]) compared to individuals classified as having had established infection at blood sample collection date.

0.00% nucleotide ambiguity cut-off:

When applying the 0.00% nucleotide ambiguity cut-off, together with a minimum sequence length requirement of 800bp, 21686/43002 (50.4%) UKHIVDRD individuals were classified as treatment naive at blood sample collection date. 3040/21686 (14.0%) of these individuals were categorised as having had a recent infection at blood sample collection date.

Re-examination of the clinical data in combination with the 0.00% nucleotide ambiguity cut-off found a largely similar pattern to that seen with the 0.17% cut-off, with treatment naive individuals categorised as having had recent infection at blood sample collection date being more likely to:

i) have higher median first and last (before treatment) $CD4^+$ cell counts (480 cells/mm^3 vs. 373 cells/mm^3, Mann-Whitney p-value < 0.0001 and 362 cells/mm^3 vs. 268 cells/mm^3, Mann-Whitney p-value < 0.0001 respectively),

ii) have a lower last viral load (before treatment) (25900 copies/ml vs. 39000 copies/ml, Mann-Whitney p-value < 0.0001),

iii) be male (Fisher's exact test p < 0.0001, OR 2.75 [CI 95: 2.28, 3.33]),

iv) be identified with the men who have sex with men (MSM) risk group (Fisher's exact test p < 0.0001, OR 1.93 [CI 95: 1.71, 2.19]), and

v) have white ethnicity (Fisher's exact test p < 0.0001, OR 2.19 [CI 95: 1.93, 2.50]) compared to individuals classified as having had established infection at blood sample collection date.

4.3.4 Relative proportions of recent and established infections linked to clusters

1. <u>New diagnosis cohort</u>

For the 0.17% nucleotide ambiguity cut-off, a greater absolute number of established infections were found to map to clusters than recent infections (366/519, 70.5% vs. 153/519, 29.5% respectively). However, individuals categorised as having had recent infection at blood sample collection date were more likely than not to be linked to a transmission cluster compared to individuals categorised as having had an established infection (p < 0.0001, OR 1.72 [CI 95: 1.31, 2.26]). Broken down by subtype this association was found to be statistically significant for subtypes B and C only (Table 4.3.).

	Patients with recent infection		Patients with established infection			
Subtype	**Linked to clusters**	**Not linked to clusters**	**Linked to clusters**	**Not linked to clusters**	**p-value**	**odds ratio [CI 95%]**
A	2	4	12	25	1.00	1.04 [0.08,8.50]
B	128	98	282	386	<0.05	1.79 [1.30,2.45]
C	17	4	48	83	<0.05	7.25 [2.20,31.36]
CRF01_AE	4	5	3	7	0.65	1.81 [0.20,18.60]
CRF02_AG	0	1	5	9	1.00	0.00 [0.00,77.91]
D	0	0	1	7	1.00	0.00 [0.00,Inf]
G	1	1	3	12	0.43	3.61 [0.04,338.34]
non_major	3	0	10	0	1.00	0.00 [0.00,Inf]
U	0	25	0	45	1.00	0.00 [0.00,Inf]
Total	155	138	364	574	<0.0001	1.77 [1.35,2.33]

Table 4.3. Proportions of new diagnosis cohort individuals with recent or established infection linked to clusters, using the 0.17% nucleotide ambiguity cut-off.

Application of the 0.00% nucleotide ambiguity cut-off found that 395/519 (76.1%) of infections mapped to clusters were established and 124/519 (23.9%) were recent infections. The relative proportions of individuals linked to clusters within each group showed a similar pattern to that seen for the 0.17% cut-off: individuals categorised as having had recent infection at blood sample collection date were more likely than not to be linked to transmission clusters compared to individuals categorised as having had an established infection ($p < 0.002$, OR 1.56 [CI 95: 1.17, 2.09]).

2. Overall UK HIV Drug Resistance Database

The relative proportions of recent and established infections linked to clusters were assessed for all 23320 treatment naive individuals in the UKHIVDRD. 9488/23320 (40.7%) treatment naive individuals were linked to clusters compared to 13832/23320 (59.3%) not linked to clusters. 3081/5975 (51.6%) recent infections were linked to clusters compared to 6407/17345 (36.9%) established infections (Fisher's exact test p-value < 0.0001, OR 1.82 (95 CI: 1.71, 1.93), indicating individuals with recent infection were more likely than not to be linked to a cluster compared to individuals with established infection for subtypes B, C, CRF02_AG, the non-major subtype subset and the unassigned subtype subset (Table 4.4a.). The remaining subtype subsets did not show any significant difference between recent infections and established infections in terms of the odds of belonging to a cluster.

A similar pattern was found when using the 0.00% nucleotide ambiguity definition of recent infection, with recent infections in subtypes B, C, CRF02_AG and the non-major subtypes being more likely than not to be linked to clusters compared to established

infections. The unassigned subtype group no longer showed a significant difference between the two categories (Table 4.4b.).

Subtype	Total treatment naive	Treatment naive patients with recent infection			Treatment naive patients with established infection			p-value	odds ratio [CI 95%]
		Total recent treatment naive	Linked to clusters	Not linked to clusters	Total established treatment naive	Linked to clusters	Not linked to clusters		
A	982	184	59	125	798	225	573	0.32	1.20 [0.83,1.72]
B	12491	3921	**2193**	**1728**	8570	**3685**	**4885**	**<0.0001**	**1.68 [1.56,1.82]**
C	5822	879	**394**	**485**	4943	**1452**	**3491**	**<0.0001**	**1.95 [1.68,2.27]**
CRF01_AE	481	140	39	101	341	69	272	0.07	1.52 [0.94,2.45]
CRF02_AG	**971**	**210**	**92**	**118**	**761**	**240**	**521**	**<0.05**	**1.69 [1.22,2.34]**
D	181	18	6	12	163	40	123	0.40	1.53 [0.44,4.77]
G	472	104	34	70	368	116	252	0.81	1.06 [0.64,1.72]
Non-major	308	80	**59**	**21**	228	**119**	**109**	**<0.05**	**2.57 [1.43,4.75]**
U	1612	439	**205**	**234**	1173	**461**	**712**	**<0.05**	**1.35 [1.08,1.70]**
Total	**23320**	**5975**	**3081**	**2894**	**17345**	**6407**	**10938**	**<0.0001**	**1.82 [1.71,1.93]**

Table 4.4a. Is the proportion of recent infections linked to clusters significantly different to the proportion of established infections linked to clusters? Table shows counts of overall numbers of treatment naive individuals in each subtype, the number of treatment naives further subdivided into infections that were recent or established at blood sample collection date (based on 0.17% nucleotide ambiguity cut-off), and then the numbers of recent and established infections further subdivided into individuals linked or not linked to clusters. Fisher's exact test p-value was derived from comparing the proportion of clustering/non-clustering individuals in the recent infection category versus those in the established infection category.

Subtype	Total treatment naive	Treatment naive patients with recent infection			Treatment naive patients with established infection			p-value	odds ratio [CI 95%]
		Total recent treatment naive	Linked to clusters	Not linked to clusters	Total established treatment naive	Linked to clusters	Not linked to clusters		
A	900	103	37	66	797	226	571	0.13	1.42 [0.89,2.22]
B	**11708**	**1969**	**1121**	**848**	**9739**	**4398**	**5341**	**<0.0001**	**1.61 [1.45,1.77]**
C	**5467**	**476**	**210**	**266**	**4991**	**1549**	**3442**	**<0.0001**	**1.75 [1.44,2.13]**
CRF01_AE	456	81	20	61	375	83	292	0.66	1.15 [0.62,2.07]
CRF02_AG	854	93	41	52	761	266	495	0.09	1.47 [0.92,2.32]
D	175	6	3	3	169	43	126	0.19	2.91 [0.38,22.55]
G	395	41	17	24	354	118	236	0.30	1.42 [0.68,2.87]
Non-major	**284**	**39**	**30**	**9**	**245**	**138**	**107**	**0.02**	**2.58 [1.13,6.44]**
U	1447	232	109	123	1215	507	708	0.15	1.24 [0.92,1.66]
Total	21686	3040	1588	1452	18646	7328	11318	<0.0001	1.69 [1.56,1.83]

Table 4.4b. Shows the same information as Table 4.4a. using a nucleotide ambiguity cut-off of 0.00% and minimum *pol* sequence length of 800bp, and shows that the significant relationships broadly hold.

The greater number of large clusters identified for subtypes B and C warranted further investigation of the effect of the size of the cluster on the proportions of recent and established infections linked to transmission clusters. Despite a greater absolute number of individuals with established infection at blood sample collection date being linked to clusters containing 10 or more individuals, individuals categorised as having had recent infection at date of blood sampling (here defined by the 0.17% nucleotide ambiguity cut-off only) were still found to be more likely to belong to a cluster than not (Table 4.5.). A similar trend existed for individuals linked to transmission pairs.

	Recent infections		Established infections			
Subtype	**Linked to clusters ≥ 10**	**Not linked to clusters ≥ 10**	**Linked to clusters ≥ 10**	**Not linked to clusters ≥ 10**	**p-value**	**odds ratio [CI 95%]**
B	808	3113	1111	7459	<0.0001	1.74 [1.57,1.93]
C	38	841	73	4870	<0.0001	3.01 [1.97,4.55]
Subtype	**Linked to transmission pair**	**Not linked to transmission pair**	**Linked to transmission pair**	**Not linked to transmission pair**	**p-value**	**odds ratio [CI 95%]**
B	496	3425	974	7596	<0.05	1.13 [1.00,1.27]
C	183	696	842	4101	<0.05	1.28 [1.06,1.54]

Table 4.5. The proportion of recent infections (here defined by the 0.17% nucleotide ambiguity cut-off) linked to clusters by cluster size: clusters involving 10 or more individuals showed a significant difference in the proportion of recent infections linked to them compared to equivalent proportion in established infections, a similar pattern existed for transmission pairs.

In order to investigate potential ascertainment bias inherent in the transmission cluster identification stage, due to incorporation of a genetic distance measurement, a less stringent bootstrapping value of ≥85% and a range of maximum intra-cluster pairwise genetic distances were used to identify transmission clusters (0.015 to 0.065), and an analysis of the proportion of recent infections versus the proportion of established infections linked to clusters was carried out. To a large extent, all statistically significant associations were found to hold across each criteria combination (Table 4.6.).

		85%					95%				
		Recent		Established			Recent		Established		
Subtype	Genetic distance	Linked	Not linked	Linked	Not linked	p-value	Linked	Not linked	Linked	Not linked	p-value
A	1.5	118	332	569	1341	0.15	99	351	482	1428	0.16
	2.5	156	294	690	1220	0.59	123	327	570	1340	0.30
	3.5	179	271	742	1168	0.75	138	312	599	1311	0.82
	4.5	190	260	754	1156	0.29	146	304	610	1300	0.87
	5.5	205	245	808	1102	0.22	146	304	610	1300	0.87
	6.5	208	242	829	1081	0.29	146	304	610	1300	0.87
B	1.5	2419	3809	5239	10857	<0.0001	1866	4362	3992	12104	<0.0001
	2.5	3130	3098	6701	9395	<0.0001	2402	3826	5052	11044	<0.0001
	3.5	3642	2586	7897	8199	<0.0001	2796	3432	5943	10153	<0.0001
	4.5	3971	2257	8646	7450	<0.0001	3087	3141	6514	9582	<0.0001
	5.5	4099	2129	8902	7194	<0.0001	3180	3048	6742	9354	<0.0001
	6.5	4126	2102	8987	7109	<0.0001	3189	3039	6756	9340	<0.0001
C	1.5	654	1208	2951	6128	0.03	570	1292	2518	6561	0.01
	2.5	795	1067	3436	5643	<0.0001	684	1178	2877	6202	<0.0001
	3.5	857	1005	3624	5455	<0.0001	730	1132	3005	6074	<0.0001
	4.5	879	983	3672	5407	<0.0001	758	1104	3048	6031	<0.0001
	5.5	891	971	3706	5373	<0.0001	765	1097	3066	6013	<0.0001
	6.5	892	970	3707	5372	<0.0001	765	1097	3066	6013	<0.0001
CRF01_AE	1.5	66	150	153	369	0.79	58	158	123	399	0.35
	2.5	79	137	179	343	0.55	63	153	130	392	0.23
	3.5	80	136	181	341	0.55	63	153	130	392	0.23
	4.5	80	136	181	341	0.55	63	153	130	392	0.23
	5.5	80	136	181	341	0.55	63	153	130	392	0.23
	6.5	80	136	181	341	0.55	63	153	130	392	0.23
CRF02_AG	1.5	120	220	452	936	0.34	113	227	391	997	0.07
	2.5	151	189	547	841	0.10	137	203	461	927	0.02
	3.5	158	182	578	810	0.11	144	196	485	903	0.01
	4.5	166	174	595	793	0.05	146	194	488	900	0.01
	5.5	167	173	596	792	0.04	146	194	488	900	0.01
	6.5	167	173	596	792	0.04	146	194	488	900	0.01
D	1.5	19	66	107	320	0.68	19	66	101	326	0.89
	2.5	21	64	129	298	0.36	21	64	119	308	0.60
	3.5	23	62	138	289	0.37	22	63	120	307	0.79
	4.5	27	58	145	282	0.80	23	62	126	301	0.70
	5.5	27	58	145	282	0.80	23	62	126	301	0.70
	6.5	27	58	145	282	0.80	23	62	126	301	0.70
G	1.5	47	135	195	486	0.52	38	144	171	510	0.28
	2.5	63	119	241	440	0.86	51	131	209	472	0.52
	3.5	72	110	266	415	0.93	56	126	223	458	0.66
	4.5	74	108	276	405	1.00	57	125	228	453	0.60
	5.5	74	108	276	405	1.00	57	125	228	453	0.60
	6.5	74	108	277	404	1.00	57	125	228	453	0.60
Non-major	1.5	49	93	136	297	0.53	46	96	128	305	0.53
	2.5	59	83	161	272	0.37	55	87	152	281	0.48
	3.5	68	74	198	235	0.70	61	81	183	250	0.92
	4.5	92	50	231	202	0.02	88	54	218	215	0.02
	5.5	95	47	236	197	0.01	89	53	223	210	0.03
	6.5	95	47	236	197	0.01	89	53	223	210	0.03

Table 4.6. Table showing how changes in bootstrap cut-off support values and maximum pairwise genetic distance cut-offs (substitutions per nucleotide) in cluster identification affects the proportion of recent (as defined by 0.17% ambiguity cut-off) and established infections found to be linked to clusters (N.B. absolute counts in the table are prior to splitting or removal of non-UK clusters). The majority of significant differences remained, with changes in bootstrap and distance cut-off making a

difference for subtype A, CRF02_AG, the non-major subtypes and the unassigned subtypes – though plateauing of the numbers of clusters found at increasing genetic distance cut-off was seen.

4.3.5 Drug resistance in the UK HIV Drug Resistance Database

All 43002 non-duplicate patient sequences were analysed for major drug resistance mutations using Sierra, the Stanford HIV Web Service submission tool (http://hivdb.stanford.edu/DR/webservices/index.html), ignoring all drug resistance mutations not listed in the WHO 2009 Drug Resistance Mutation List (Bennett, Camacho et al. 2009). 1924/23320 (8.3%) treatment naive individuals showed evidence of one or more transmitted drug resistance mutations acquired from individuals harbouring drug resistant virus (Table 4.7.). Both the 0.17% and 0.00% nucleotide ambiguity cut-offs for identifying recent infection were imposed to investigate the proportion of transmitted drug resistance (TDR) in recent infections compared to established infections. It was found that there was proportionally greater TDR in recent infections when using both cut-offs (Table 4.8.).

	No. of patients across all subtypes (n=23320)	No. of subtype B patients (n=12491)
Overall DR (in any class)	1924 (8.3%)	1360 (10.9%)
NNRTI resistance	1924 (8.3%)	1360 (10.9%)
NRTI resistance	788 (3.4%)	467 (3.7%)
PI resistance	404 (1.7%)	258 (2.1%)
IN resistance	0 (0.0%)	0 (0.0%)
Resistance to ≥2 classes	1121 (4.8%)	670 (5.4%)
Overall DR in recent infections	573 (2.5%)	464 (3.7%)
Overall DR in established infections	1351 (5.8%)	896 (7.2%)
Overall DR in patients linked to clusters	897 (3.8%)	706 (5.7%)
Overall DR in patients not linked to clusters	1027 (4.4%)	654 (5.2%)

Table 4.7. Table showing numbers of treatment naive individuals with drug resistance mutations for all subtypes and for subtype B (the subtype expected to contain the greatest number of individuals with drug resistance), categorised by drug class, infection status, and whether or not they are linked to clusters (DR = drug resistance, NNRTI = non-nucleoside reverse transcriptase inhibitor, NRTI = nucleoside reverse transcriptase inhibitor, PI = protease inhibitor, IN = integrase inhibitor).

0.17%

	TDR	no TDR	Total
Recent	566	5409	5975
Established	1358	15987	17345
Total	1924	21396	23320

$X^2 = 15.64$ p-value < 0.0001

0.00%

	TDR	no TDR	Total
Recent	290	2750	3040
Established	1510	17136	18646
Total	1800	19886	21686

$X^2 = 6.95$ p-value = 0.0084

Table 4.8. Is the proportion of recent infections harbouring transmitted drug resistance (TDR) greater than the proportion of established infections harbouring TDR? Upper table shows counts for 0.17% nucleotide ambiguity cut-off, bottom table shows counts for 0.00% nucleotide ambiguity cut-off. In both instances there is a significant difference in proportions (Pearson's Chi-squared test with Yates' continuity correction).

4.3.6 Co-clustering of treatment naive drug susceptible and treatment experienced drug resistant individuals

To assess the potential for treatment naive individuals categorised as having had a recent infection at blood sample collection date to be infected with a drug resistant virus from anti-retroviral treatment (ART) experienced individuals failing treatment, an assortativity coefficient was calculated to gauge the extent of co-clustering between these patient categories. For each subtype subset, a matrix of pairwise connections within and between treatment naive individuals with recent infection (using both the 0.17% and 0.00% ambiguity cut-offs) but no evidence of drug resistant virus, and treatment experienced individuals with drug resistant virus (regardless of infection stage) was set up, and an assortativity coefficient calculated as described elsewhere (Newman 2003) and previously used in a similar context (Volz, Koopman et al. 2012) (Table 4.9a.). An assortativity coefficient of 0 indicates no assortative mixing (i.e. there are as many connections between individuals from different categories as there are between individuals from the same category); whereas a coefficient of 1 indicates perfect assortative mixing (i.e. connections are only between individuals of the same category).

When the 0.17% ambiguity cut-off was used to identify individuals with recent infections, the overall assortative coefficient for the clusters was 0.45. By subtype subset, all subtypes except D had a positive coefficient value – though caution should be employed when dealing with small numbers of infections. These results, together with the initial analysis above, indicate a strong likelihood of drug susceptible treatment naive individuals with recent infection to co-cluster with each other, and likewise drug resistant treatment experienced individuals to co-cluster with each other. However, there is clearly opportunity for transmission of drug resistant virus between the categories - particularly for subtype B, where most drug resistance was found, which had more connections between the two groups

than within the treatment experienced individuals with drug resistant virus (485 vs. 401 respectively).

When the 0.00% ambiguity cut-off was applied the assortativity coefficient remained positive for all subtypes (with the exception of subtype D again), but was smaller in value for all subtype except B, CRF02_AG and the non-major subtype subset, where the value increased (Table 4.9b.).

Subtype	Recent, no DR ↔ Recent, no DR	Recent, no DR ↔ Experienced, DR	Experienced, DR ↔ Experienced, DR	Assortativity coefficient
A	12	3	24	0.69
B	5982	485	401	0.38
C	173	51	125	0.48
CRF01_AE	7	3	7	0.40
CRF02_AG	36	2	6	0.70
D	0	1	5	-0.17
G	15	0	3	1.00
Non-major	419	4	3	0.42
U	138	18	49	0.62
Total	6782	567	623	0.45

Table 4.9a. Counts of pairwise connections between treatment naive drug susceptible individuals with recent infection and treatment experienced drug resistant individuals as categorised using the 0.17% ambiguity cut-off, where counts are of pairwise connection within the same cluster. Assortativity coefficients are calculated from the matrices of these pairwise counts, with values from > 0 to 1 indicating more connections between individuals of the same category than between different patient categories.

Subtype	Recent, no DR ↔ Recent, no DR	Recent, no DR ↔ Experienced, DR	Experienced, DR ↔ Experienced, DR	Assortativity coefficient
A	2	3	24	0.29
B	1961	252	401	0.50
C	42	31	125	0.38
CRF01_AE	2	3	7	0.10
CRF02_AG	12	0	6	1.00
D	0	0	5	-
G	1	0	3	1.00
Non-major	69	0	3	1.00
U	49	13	49	0.58
Total	2138	302	623	0.55

Table 4.9b. Counts of pairwise connections between treatment naive drug susceptible individuals with recent infection and treatment experienced drug resistant individuals as categorised using the 0.00% ambiguity cut-off, where counts are of pairwise connection within the same cluster.

4.3.7 Modelling recent infection dynamics within clusters

A model of intra-cluster transmission dynamics was developed to gauge the extent to which the initial stages of HIV infection contribute to onward transmission in the context of transmission clusters.

In an attempt to test the sensitivity of the model results to the length of the recent infection phase, the definition of the recent phase of infection was altered in steps of 20 days, from 20 to 400 days in length. This was carried out on three different cluster size categories: clusters containing a minimum of either 2, 3 or 10 recent infections, in order to investigate dynamics in clusters of different sizes (N.B. although cluster size does not correspond directly with the number of recent infections in a cluster, there is a strong linear correlation, with an R^2 value of 0.84 (data not shown)). For all analyses, a count of the number of possible transmission events that could have taken place within the cluster was made (only counting individuals categorised as recent infections), together with a measure of the proportion of these potential transmissions that would have occurred whilst the individual in question was still in the recent phase stage of infection. Results for the UK phylogeny based clusters and the reconstructed clusters with randomised dates of seroconversion were then compared for each cluster size and recent phase window size, using Fisher's exact test to determine if the counts deviated significantly between the two cluster types.

Results of the model run with the minimum number of recent infections required per cluster set to 2 revealed that for subtype B, significantly more potential recent phase transmissions took place in the reconstructed clusters with randomised infection dates compared to UK phylogeny based clusters the longer the recent infection phase windows (i.e. for the 240, and 280-400 day windows) (Table 4.10a). In contrast, when the minimum number of recent infections per cluster was set to 3, recent infection window lengths of 20 and 40 days were found to result in a significantly higher number of potential recent phase

transmissions in the UK phylogeny based clusters compared to the reconstructed, randomised date clusters. When the minimum number of recent infections per cluster was set to 10, a recent infection window length of 20 days was found to result in a significantly higher number of potential recent phase transmissions in the actual clusters compared to the reconstructed, random date clusters. Subtype C did not reveal any significant differences across cluster size or window length (Table 4.10a.), but the number of eligible clusters available when the minimum number of recent infections per cluster was set to 10 was just 1.

When the model was re-run using the 0.00% ambiguity cut-off to categorise recent infections, the overall number of eligible clusters was reduced across all size categories for subtype B and Subtype C. Only subtype B revealed a significant difference in numbers of recent phase infections between the actual UK phylogeny clusters and the reconstructed, random date clusters. This was for clusters with a minimum of 3 recent infections, with a recent phase window length of 20 days (Table 4.10b.).

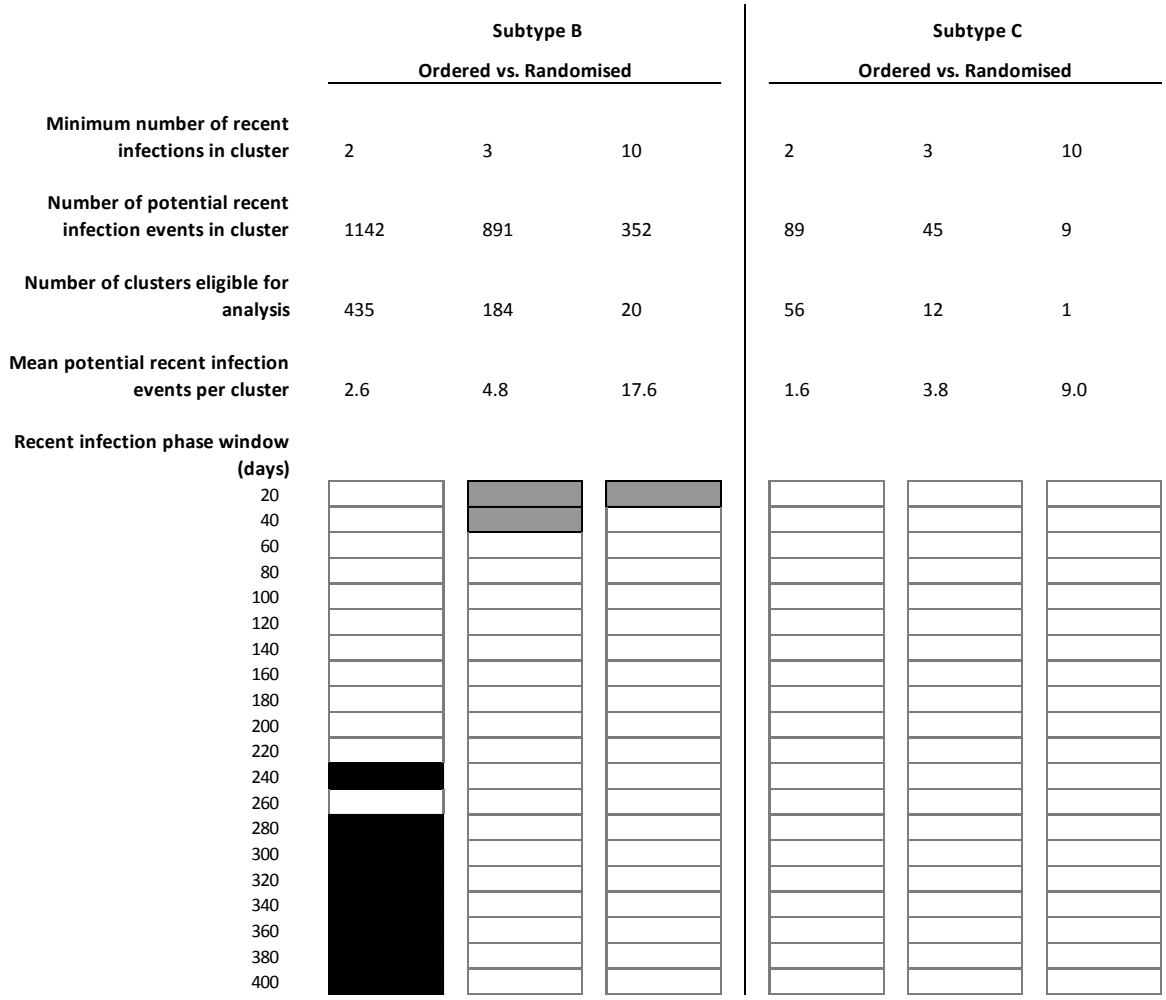


Table 4.10a. Transmission model results for subtypes B and C for three categories of minimum number of infections required per cluster. Black blocks indicate significantly higher numbers of potential recent phase transmissions in the constructed clusters compared to the UK phylogeny based clusters; grey blocks indicate significantly higher numbers of potential recent phase transmissions in the UK phylogeny based clusters compared to the constructed clusters; clear blocks indicate no significant difference. For subtype B, greater numbers of recent phase infections at short window lengths in comparison to the randomised infection date scenario implies a higher degree of infections overlap at this window size compared to the randomised date scenario, in turn suggesting recent infections (20 to 40 days post-seroconversion) may tend to infect other recent infections whilst still in the very early stages of infection.

	Subtype B Ordered vs. Randomised			Subtype C Ordered vs. Randomised		
Minimum number of recent infections in cluster	2	3	10	2	3	10
Number of potential recent infection events in cluster	443	324	109	31	14	10
Number of clusters eligible for analysis	191	72	6	22	5	0
Mean potential recent infection events per cluster	2.3	4.5	18.2	1.4	2.8	-
Recent infection phase window (days)						
20						
40						
60						
80						
100						
120						
140						
160						
180						
200						
220						
240						
260						
280						
300						
320						
340						
360						
380						
400						

Table 4.10b. The model was re-run using a 0.00% ambiguity cut-off to categorise recent infection. Grey blocks indicate significantly higher numbers of potential recent phase transmissions in the UK phylogeny based clusters compared to the constructed clusters; clear blocks indicate no significant difference. For subtype B, greater numbers of recent phase infections at a 20 day recent phase window length in comparison to the randomised infection date scenario implies a higher degree of infections overlap at this window size compared to the randomised date scenario, in turn suggesting recent infections may tend to infect other recent infections whilst still in the very early stages of infection.

4.3.8 Counting recent phase transmissions using nucleotide ambiguity and pairwise genetic distance

An investigation into combining the nucleotide ambiguity classifier with a minimum pairwise genetic distance cut-off to identify putative recent phase infections was carried out across subtypes, on three different cluster size categories: pairs, clusters containing 3 to 9 individuals, and clusters containing 10 or more individuals. Briefly, for each cluster analysed, all pairwise relationships falling below a cut-off of 0.03 nucleotide substitutions were then checked to see if they involved individuals categorised as recent infections on the basis of nucleotide ambiguity, and if so, whether or not their 125 day recent phase infection windows overlapped. If these three conditions were met, the pair was deemed to be a potential recent phase transmission event (though it is important to note that the possibility that a third person, not captured by the UKHIVDRD, infected both individuals at a similar time point cannot be ruled out by this method).

Using the 0.17% ambiguity cut-off produced a greater overall number of putative recent phase infections for all subtypes and all cluster size categories, compared to using the more stringent 0.00% ambiguity cut-off (Table 4.11.). Subtype B showed a significant increase in the proportion of total potential cluster transmissions (determined by subtracting 1 from the overall cluster size, this time including both recent and established infections) coming from the recent phase of infection when increasing the cluster size category (i.e. transmission pairs vs. clusters containing 3 to 9 individuals, and clusters containing 3 to 9 individuals vs. clusters containing 10 individuals or more) (Fisher's exact test, p-value < 0.05). For subtype C, there was a significant increase in the proportion of total potential cluster transmissions coming from the recent phase of infection when moving from clusters containing 3 to 9 individuals to clusters containing 10 individuals or more, for both 0.00% and 0.17% ambiguity cut-offs. The non-major subtype subset had significant increases in the

proportion of total potential cluster transmissions coming from the recent phase of infection moving from pairs to clusters containing 3 to 9 individuals, and from clusters containing 3 to 9 individuals to clusters containing 10 individuals or more, but this was for the 0.17% ambiguity cut-off only. Finally, the unassigned subtype subset showed a significant increase in the proportion of total potential cluster transmissions coming from the recent phase of infection when moving from pairs to clusters containing 3 to 9 individuals for the 0.17% ambiguity cut-off only. The remaining subtypes did not reveal any significant differences in the proportion of total potential cluster transmissions coming from the recent phase of infection across cluster size categories.

Cluster size	B			C		
	potential transmissions (cluster size -1)	recent phase transmissions (% total)		potential transmissions (cluster size -1)	recent phase transmissions (% total)	
		0.00%	0.17%		0.00%	0.17%
pairs	1400	**8 (0.6)***	**21 (1.5)***	1059	4 (0.4)	12 (1.1)
3 to 9	3130	**39 (1.2)*†**	**126 (4.0)*†**	1118	**6 (0.5)†**	**21 (1.9)†**
10 up	2589	**214 (8.3)†**	**468 (18.1)†**	141	**8 (5.7)†**	**18 (12.8)†**

Cluster size	Non_major			U		
	potential transmissions (cluster size -1)	recent phase transmissions (% total)		potential transmissions (cluster size -1)	recent phase transmissions (% total)	
		0.00%	0.17%		0.00%	0.17%
pairs	58	0 (0.0)	**0 (0.0)***	266	2 (0.8)	**5 (1.9)***
3 to 9	78	3 (3.8)	**10 (12.8)*†**	423	12 (2.8)	**24 (5.7)***
10 up	83	9 (10.8)	**24 (28.9)†**	115	2 (1.7)	7 (6.1)

Table 4.11. Comparing the proportion of putative recent phase infections between cluster size categories and the two ambiguity cut-off levels for each subtype. Only subtypes revealing a significant difference between proportions (Fisher's exact test, $p < 0.05$) are shown (in bold). * indicates significant difference when comparing transmission pairs to clusters containing 3 to 9 individuals, and † indicates significant difference when comparing clusters containing 3 to 9 individuals to clusters containing 10 individuals or more.

4.4 Discussion

We investigated the role of recent infection in transmission clusters within the UK HIV-1 epidemic at two depths of focus: the first, based on a cohort of individuals newly diagnosed with HIV-1, with well-defined laboratory-based evidence of seroconversion, offered a greater degree of confidence in terms of the categorisation of individuals into those who had a recent infection at diagnosis and those who had an established infection, but with a concomitant limitation on the number of individuals available. The second level of focus was wider in scope, but necessarily less in-depth; representing a trade-off in terms of the stringency of recent/established infection categorization and a more comprehensive coverage of the overall UK HIV-1 epidemic.

The new diagnosis cohort and the UKHIVDRD were found to have similar proportions of individuals categorised as recent infections at blood sample collection date when the 0.17% nucleotide ambiguity cut-off was applied, 381/1570 (24.3%) and 5975/23320 (25.6%) respectively. Although the ambiguity classifier was incorporated into the new diagnosis cohort categorisation of recent and established infection, only in 279/1570 (17.8%) cases was it used in isolation, where there was no other evidence to aid classification, and when these patients were excluded, the proportion of recents was only slightly lower (285/1291 (22.1%)). The similarity of the proportions for the new diagnosis cohort and the UKHIVDRD lends support to the validity of the nucleotide ambiguity analysis methodology. When the 0.00% nucleotide ambiguity cut-off was applied, the proportion of recent infections in the two cohorts diverged somewhat, with 20.2% and 14.0% of infections being categorised as recent in the new diagnosis cohort and UKHIVDRD respectively. It is difficult to compare these figures with results from other studies, as location, demographics in terms of ethnicity and risk group, recent infection ascertainment and definition and other factors differ widely between studies – however both the upper range of 25.6% and the lower

range of 14.0% have support in the literature. Using various combinations of serological evidence to identify recent infection, 24.9% of 7902 French patients diagnosed in 2003-2005 were classified as having recent infection at sample collection date (Semaille, Barin et al. 2007); Fisher et al. found that 19% of 859 MSM attending a UK clinic between 2000 and 2006 had recent infections (Fisher, Pao et al. 2010); a similar percentage of 20% was found by Yerly et al. in a Swiss cohort of 637 individuals diagnosed between 2000 and 2008 (Yerly, Junier et al. 2009). Whilst the proportion of infections categorised as recent in the UKHIVDRD when applying the 0.00% nucleotide seems substantially different to these previous studies, it matches well with the proportion of recent infections estimated from Recent Infection Testing Algorithm (RITA) results obtained for 3070/5970 (51%) people newly diagnosed in England, Wales and Northern Ireland in 2011 (Health Protection Agency 2012). The results of that study suggested that 16% of newly diagnosed subjects had been infected in the previous 4 to 6 months.

It is also reassuring to observe that UKHIVDRD individuals identified as recent infections using the nucleotide ambiguity classifier displayed statistically significant differences in their demographic and clinical profile compared to those categorised as having had an established infection at blood sample collection, and that these differences make sense in the context of other studies looking at clinical and epidemiological differences between individuals with recent versus established HIV-1 infections, e.g. higher baseline $CD4^+$ cell count, white ethnicity and MSM risk group (Manavi, McMillan et al. 2004; Girardi, Sabin et al. 2007).

This study found that a disproportionate number of individuals with recent infections were linked to clusters compared to individuals with established infections, for both the new diagnosis cohort and the overall UKHIVDRD. This was predominantly driven by subtypes B and C for the new diagnosis cohort, but expanded to include subtypes B, C, CRF02_AG and

the non-major subtype subset for the overall UKHIVDRD. The extent to which transmission clusters may have enabled transmission of drug resistant virus from treatment experienced individuals, failing therapy, to uninfected individuals was investigated using an assortativity coefficient to measure the degree of mixing between treatment naive individuals with recent infection but without drug resistant virus, and treatment experienced individuals (regardless of infection stage) with drug resistant virus. A relatively low level of mixing seemed to occur between these two categories, but there is clear scope for transmission of drug resistant virus from treatment experienced individuals to treatment naive individuals within transmission clusters, particularly in subtype B.

As highlighted in the introduction, other studies have attempted to use phylogenetic approaches to assess the extent to which individuals in the initial stages of HIV infection might go on to infect others, and have found broadly similar patterns of disproportionate linkage of individuals with recent infection to clusters (Brenner, Roger et al. 2007; Fisher, Pao et al. 2010). However, without knowing when transmissions took place within a cluster, it is likely to be misleading to assume that linkage of a recent infection to a cluster means that there was a transmission from the recent infection phase. Recent infection is by definition transient, and an individual may have their virus sampled soon after infection, but may not transmit their infection until it is more established (or not at all) (Brown, Gifford et al. 2009). The two methods developed in this study were attempts to address this issue by aiming to construct proxy measurements of the actual number of transmissions from the recent infection phase, which may enable informative comparisons between different demographic sub-structures of the epidemic (e.g. MSM vs. heterosexual risk group, transmission pairs vs. larger clusters).

The first method used UK-based phylogenies to identify transmission clusters to enable a count of the number of recent infections overlapping with each other within each

cluster to be made, and then compared this to clusters of equivalent size, containing an identical proportion of recent infections, but with randomised seroconversion dates. Results showed that for subtype B, when the period of recent infection was defined as 20 or 40 days, there was a significantly greater degree of overlap of recent infections in UK phylogeny clusters compared to equivalent clusters with randomised seroconversion dates than might be expected by chance within the ≥3 recent infections or ≥10 recent infections categories. This suggests that in larger clusters there is a greater degree of temporal overlap of infections in the periods shortly after seroconversion, compared to transmission pairs. This finding lends support to the hypothesis that clusters of infections are driven by individuals still in the initial stages of infection, who are likely to have higher viral titres in their blood and genital compartments, and may be unaware of their infection status. The fact that a similar result is not observed in subtype C may reflect different dynamics at play in a sub-epidemic made up predominantly of heterosexual transmission (Tatt, Barlow et al. 2004), and could mean that individuals in this sub-epidemic are less likely to transmit whilst in the recent stage of their infection, or simply that large clusters are less likely to form. One possible strategy to address the latter hypothesis may be to perform a randomised sub-sampling of subtype B clusters, and within those clusters a randomised sub-sampling of infections, to more closely match the numbers and size of the subtype C clusters, and to repeat the phylogenetic analysis and application of the model. If the pattern for subtype B clusters is still found at this reduced sampling density, then it may suggest that the differences between the subtypes is more to do with real differences in transmission dynamics, if the result is lost at this reduced sampling density, it points to the lower number of clusters and individuals with subtype C being the primary cause of the difference in results.

For subtype B clusters containing at least 2 recent infections, it was found that higher numbers of recent phase infections occurred in the UK phylogeny base clusters compared to

the randomised infection date clusters when the period of recent infection was defined as 240 days or greater. This at first unintuitive finding may still fit with the concept of temporal grouping of infections. Infections in the randomised infection date clusters should have, by definition, a random distribution across the time span of the cluster, and therefore an increase in the recent infection period should increase the chances that adjacent recent infections will overlap with each other, particularly if the recent infection period becomes large relative to the overall time span of the cluster. However, if there is already a higher degree of temporal grouping of recent infections in the UK phylogeny based clusters, then there may not be the same kind of overall uniform spread of infections across the cluster time span, but instead a number of islands of infections, which may not overlap as quickly with other islands as the infection window length is increased (Fig 4.2.).

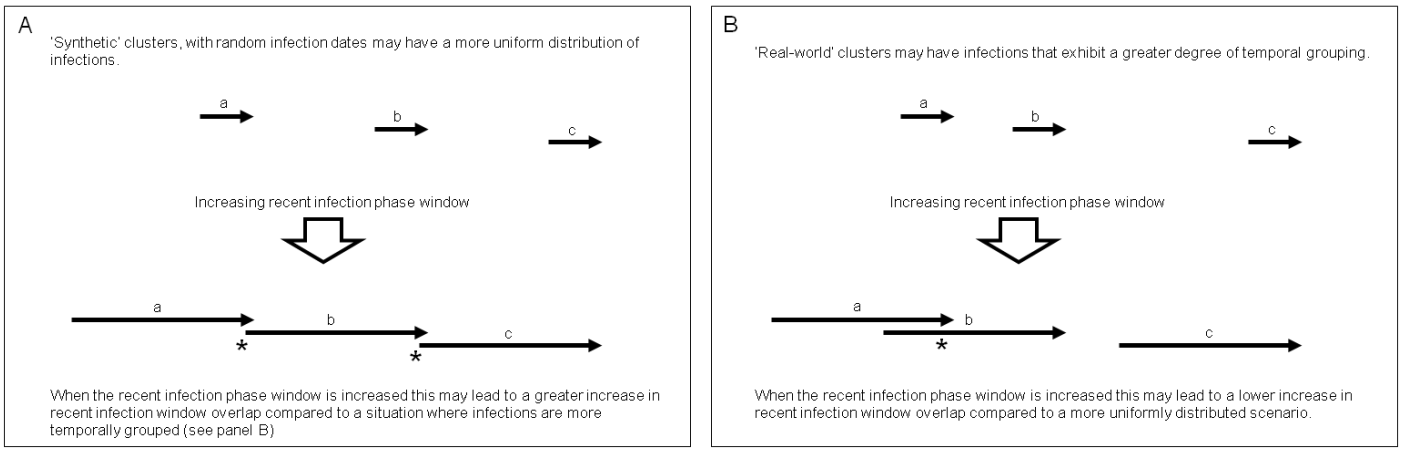


Figure 4.2. A schematic outlining why subtype B clusters with randomised seroconversion (and hence infection) dates might exhibit greater numbers of recent phase transmission events at larger recent infection phase window settings compared to UK phylogeny based clusters, * indicates a recent phase transmission event.

Overall, this model is a preliminary proof-of-concept approach, and an investigation into possible ways in which application of the nucleotide ambiguity classifier may allow a greater insight into the role of recent infection in the formation of transmission chains within the UK HIV-1 epidemic. Further work is required to investigate the sensitivity of the model, and the ability of the model to detect true temporal clustering of infections, possibly by constructing clusters with pre-determined degrees of recent infection overlap, and testing how the model behaves.

The second method employed to gauge the number of recent phase infections in clusters used a combination of recent infection overlap and a pairwise genetic distance cut-off. Because the method cannot exclude the possibility of third-party individuals not captured by the UKHIVDRD playing a role in transmission events in clusters, the method produces a proxy for the actual number of recent infection phase transmission events that have taken place in each cluster. The method found that for subtypes B, C, the non-major subtype subset, and the unassigned subtype subset, there were increases in the proportion of putative recent phase transmissions between the cluster size categories. For subtype B this was true across all three cluster size categories and at both the 0.17% and 0.00% nucleotide ambiguity cut-off, and suggests that for the subtype B epidemic in particular, there is robust support for the hypothesis that clusters of HIV-1 infections may be driven by a high proportion of transmissions from individuals in the recent phase of their infection.

It is important to acknowledge potential limitations to the methodology employed in this study. The first limitation is that the UKHIVDRD may not be representative of the overall HIV-1 epidemic in the UK. If individuals from particular risk groups, ethnicities, and socio-economic conditions are more or less likely to receive late HIV diagnoses (Manavi, McMillan et al. 2004; Boyd, Murad et al. 2005; Girardi, Sabin et al. 2007), then this may lead to the UKHIVDRD being skewed by a particular dominating risk group and/or ethnicity (in

this case white MSMs). There may also be ascertainment bias inherent in the transmission cluster identification stage, which by its very process usually incorporates a genetic distance measurement that may tend to favour identification of clusters containing individuals whose viruses have had less time to genetically diverge from each other, thereby potentially inflating the numbers of recent infections linked to transmission clusters (Volz, Koopman et al. 2012). To investigate this second possibility, less stringent combinations of bootstrap value and maximum intra-cluster pairwise genetic distances were used to identify transmission clusters. Analysis of the proportion of recent infections versus the proportion of established infections linked to clusters found that to a large extent, all statistically significant associations were found to hold across each criteria combination, indicating that the results of the analysis, particularly for subtypes B and C, are relatively robust to ascertainment bias introduced by the cluster identification methodology. Focussing on the methodologies used to assess the degree of onward transmission from the recent infection stage within clusters, it is clear that the models used necessarily represent an oversimplification of the true dynamics at play within actual transmission chains, and several caveats must be acknowledged: firstly, a caveat which applies not just to this study, but to all approaches based upon phylogenies sampled from a population, is that not all members of a transmission chain are likely to have been captured by the UKHIVDRD: some clusters may represent everyone from the actual local chain of infection, but many will not. If this difference in coverage is strongly linked to risk group, ethnic or socio-economic factors then this may lead to ascertainment biases that need to be addressed by the analysis. A closely related issue is the influence of contact tracing on the identification of transmission clusters. In the context of HIV, contact tracing primarily involves the identification of current and previous sexual partners of the infected individual who may have been unaware of exposure to the virus, and may wish to be screened for infection. Intuitively, it seems possible that for patients diagnosed with recent infection this

may lead to successful identification and subsequent testing of a greater proportion of current and previous partners compared to those diagnosed with established infections, due to the more limited timeframe within which to recall and identify these individuals. If infected by the initial HIV-1 positive individual with recent infection, these current and previous partners will necessarily have recent infections. These partners may then identify current and previous partners in turn, leading to a limited positive feedback phenomenon which may inflate the numbers of recent infections linked to clusters relative to established infections. Whilst the impact of clustering on the effectiveness of contact tracing has been investigated to some extent (Eames and Keeling 2003; House and Keeling 2010), there is a paucity of research into potential recent infection clustering bias introduced through contact tracing. The effectiveness of contact tracing in identifying infected individuals in various settings has been examined, and a systematic review of some of these studies suggest that perhaps 1-8% of individuals identified as potentially exposed to infection and not identified by other means were diagnosed with HIV through partner notification (Hogben, McNally et al. 2007). Whilst these studies did not focus on the proportion of recent infections identified through contact tracing, it may be possible to model this aspect using estimates of the incidence of recent infection taken from random testing studies in order to assess the extent to which the proportion of recent infections linked to transmission clusters may be over-inflated relative to established infections through differences in contact tracing efficiency.

In conclusion, the results of this study suggest that individuals categorised as having had a recent infection at sample date are disproportionately linked to transmission clusters in the UK epidemic, and that this effect may be compartmentalised into specific subtypes. Furthermore, there is evidence supporting a disproportionate role for onward infection from the recent stage of infection within individuals harbouring subtype B virus. There is scope for transmission of drug resistance from treatment experienced individuals to treatment naive

individuals within clusters due to individuals from both categories mixing within transmission networks, and this will require ongoing monitoring. Finally, incorporating consensus sequence based nucleotide ambiguity classifiers of HIV infection age into methodologies attempting to assess transmission dynamics of HIV epidemics – particularly when combined with pairwise genetic distance measurements – may present an effective way of investigating the extent to which individuals in different stages of infection contribute to HIV-1 epidemics, and could therefore open up investigations of this type in other national epidemics, where such sequence repository resources are available.

Chapter Five

HIV-1 deep-sequencing subpopulation dynamics in recent infection and over time

5 HIV-1 deep-sequencing subpopulation dynamics in recent infection and over time

Abstract

There is support for the concept of a genetic bottleneck operating on HIV-1 transmission during sexual contact. Such a bottleneck may serve to reduce the complexity of the virus that is transferred to the uninfected individual, possibly to the extent that a single or few viruses establish infection in the recipient. This initially homogeneous viral population then increases in diversity due to error prone replication and host immune pressure. This chapter investigates a cohort of individuals with recent infection using a deep-sequencing approach in order to ascertain the viral dynamics at play at this stage.

36 individuals, 9 of whom had multiple pre-treatment time points available for study, had their virus deep-sequenced. A read clustering approach was used to assess the extent to which the viruses developed subpopulation structure within the host due to immune pressure. At the earliest sampled time point, 30.6% patients had evidence of >1 subpopulation in *pol* and 47.2% patients had evidence of >1 subpopulation in *env*. Individuals with multiple time points displayed increases in *pol* and or *env* subpopulations over time in the majority of cases, with a few individuals retaining a single subpopulation over all time points.

The study revealed a complex mixture of HIV-1 dynamics in early infection in a relatively small cohort. The read clustering approach offered a useful perspective on how HIV-1 explores evolutionary pathways within its host; exploration that has important implications for vaccine design and antiretroviral drug treatment.

5.1 Introduction

Evidence for a genetic bottleneck operating on HIV-1 during sexual transmission was published over 20 years ago (Zhang, MacKenzie et al. 1993; Zhu, Mo et al. 1993), and has more recently gained further support (Keele, Giorgi et al. 2008; Salazar-Gonzalez, Salazar et al. 2009; Fischer, Ganusov et al. 2010). If such a bottleneck exists, it has implications for vaccine and microbicide design (McMichael, Borrow et al. 2009; Haase 2010), and may open up the possibility of using limited viral quasi-species diversity to identify recent infections (Kouyos, von Wyl et al. 2011; Poon, McGovern et al. 2011; Ragonnet-Cronin, Aris-Brosou et al. 2012). Such a bottleneck may operate through a combination of i) high levels of virion blockage at the mucosal barrier (Keele and Estes 2011), ii) stochastic attenuation of viruses due to errors introduced during replication either by the virus reverse transcriptase (RT) or innate antiretroviral factors such as the APOBEC3G cytidine deaminase (Finzi, Plaeger et al. 2006; McMichael, Borrow et al. 2009), iii) selection of particular V3 loop configurations of the envelope glycoprotein during the very early stages of infection prior to seroconversion (Zhang, MacKenzie et al. 1993) or iv) selection through differences in cell tropism between virions (Gray, Sterjovski et al. 2005). Host immune pressure begins to act upon the founding virus early after transmission, primarily through $CD8^+$ T cells, and leads to genetic diversification across the genome (Zhang, MacKenzie et al. 1993; Fischer, Ganusov et al. 2010; Henn, Boutwell et al. 2012). Some mutations may enable escape from host immune pressure whilst reducing the replicative capacity or some other aspect of viral fitness (e.g. transmissibility), and thus additional compensatory mutations may emerge and become linked. These selective mechanisms generate viral subpopulations that exist at various and fluctuating levels within an individual depending on subsequent immune and drug pressure (Huang, Gamarnik et al. 2003; Bordería, Lorenzo-Redondo et al. 2010). Some subpopulations may become the dominant quasi-species within the patient over time or become fixed at a

non-dominant minority level (Wilke 2005); some subpopulations may also become dominant in one compartment (e.g. genital) but not another (e.g. blood) if immune pressure and target cell populations differ (Haggerty and Stevenson 1991; Rozera, Abbate et al. 2009). An understanding of intra-patient quasi-species dynamics may lead to further insights into the role of host immune pressure on viral evolution, providing further clues to vaccine and microbicide development; whereas an understanding of how these subpopulations are transmitted between individuals may enable a greater understanding of the properties of HIV-1 epidemics. The phylogenetic approaches used to investigate transmission dynamics of epidemics may be confounded by lack of sufficient refinement to take into account preferential transmission of minority subpopulations or preferential outgrowth of minority subpopulations transmitted alongside major subpopulations. Directly pertinent to this latter point is evidence indicating that caution may be required when applying the assumption of single virion founded infections linearly increasing in genetic diversity over time (Li, Bar et al. 2010; Redd, Mullis et al. 2012; Wagner, Pacold et al. 2013). A donor who has had a dual infection may transmit virions containing highly diverse viruses to the recipient, or an individual may be superinfected with a HIV-1 strain very different to the initial infecting strain, situations which may lead to complex intra-patient subpopulation dynamics. These dynamics may be partially obscured by observing genetic variants on an individual nucleotide level, in contrast to a more sophisticated treatment of viral variants as being linked with each other, i.e. subpopulations of linked variants (Baccam, Thompson et al. 2001; Domingo, Sheldon et al. 2012). These subpopulations may represent species that have evolved away from the founding virus through a fitness landscape requiring multiple compensatory mutations, or may allow identification of species of different subtype to the major species, indicating a history of co- or superinfection in the current host or somewhere in the preceding transmission chain.

This study aimed to tackle some of the above issues by carrying out deep-sequencing of 36 patients with early HIV-1 infection in both *pol* and *env*. The study assessed the complexity of the viral quasi-species within patient plasma samples, with a view to detecting the extent to which the infections are likely to have been founded by single virions. Nine patients had multiple time points available, allowing insight into how the quasi-species changes over time. A putative transmission cluster detected using traditional Sanger *pol* sequence was also investigated using deep-sequencing, in an attempt to uncover complexities of quasi-species transmission dynamics not revealed through consensus sequencing.

5.2 Methods

5.2.1 Patient samples

Samples were obtained from 36 anonymised patients newly diagnosed with HIV-1 infection between 2004 and 2009 at the HIV service of the Royal Free London NHS Foundation Trust. Plasma samples were obtained as part of routine procedures and clinical data such as viral load and $CD4^+$ cell count were collected. A guanidine based HIV antibody avidity assay was also performed at diagnosis to assess likely length of HIV infection (Chawla, Murphy et al. 2007).

5.2.2 Sample preparation

Multiple samples subsequent to diagnosis were available for 9/36 patients, with the remaining 27 patients having one time point only. All samples were subjected to 454 pyrosequencing as previously described (section 2.10).

5.2.3 Subpopulation identification

Fastq reads were quality filtered using a maximum expected error cut-off of 0.5, as implemented in the UPARSE pipeline (Edgar 2013). A minimum length open-reading frame filter of 390bp for the *pol* amplicon and 380bp for the *env* amplicon was imposed to retain only reads likely to encode functional gene products (i.e. avoiding reads with artificial homopolymer insertions). Forward and reverse reads were combined and aligned using Muscle (Edgar 2004). Any columns in the alignment containing gaps were then removed, and reads clustered into operational taxonomic units (henceforth referred to as subpopulations) as per the UPARSE pipeline, using a 1% genetic distance cut-off. Reads were then mapped to

subpopulations using a ≥99% identity cut-off (Fig 5.1). Sensitivity analysis was carried out to observe whether or not altering the distance cut-off would change the number of subpopulations identified within each sample: a more generous cut-off of 3% was substituted into the original pipeline, with all other parameters remaining unaltered. Read counts for subpopulations identified using the 1% cut-off were subjected to a margin of error correction using a 99% confidence interval (Lohr 2009). A cut-off based upon the lower boundary margin of error corrected percentage was used to decide whether the subpopulation was present at abundance greater than the chosen read abundance cut-off of 1% (section 2.12). Once putative subpopulations had been identified in this manner they were subjected to a further filtering step using BLAST and genetic distance to identify minority subpopulations that matched to within a <1% genetic distance of another subpopulation. If the matching subpopulation was not seen in additional time points from the same patient, it was excluded from the analysis on the basis that it could be the result of MID tag switching (Carlsen, Aas et al. 2012). Subtyping of each subpopulation within each sample was carried out using COMET (Pineda-Peña, Faria et al. 2013), and confirmed by the Los Alamos Recombinant Identification Program web tool (Siepel, Halpern et al. 1995).

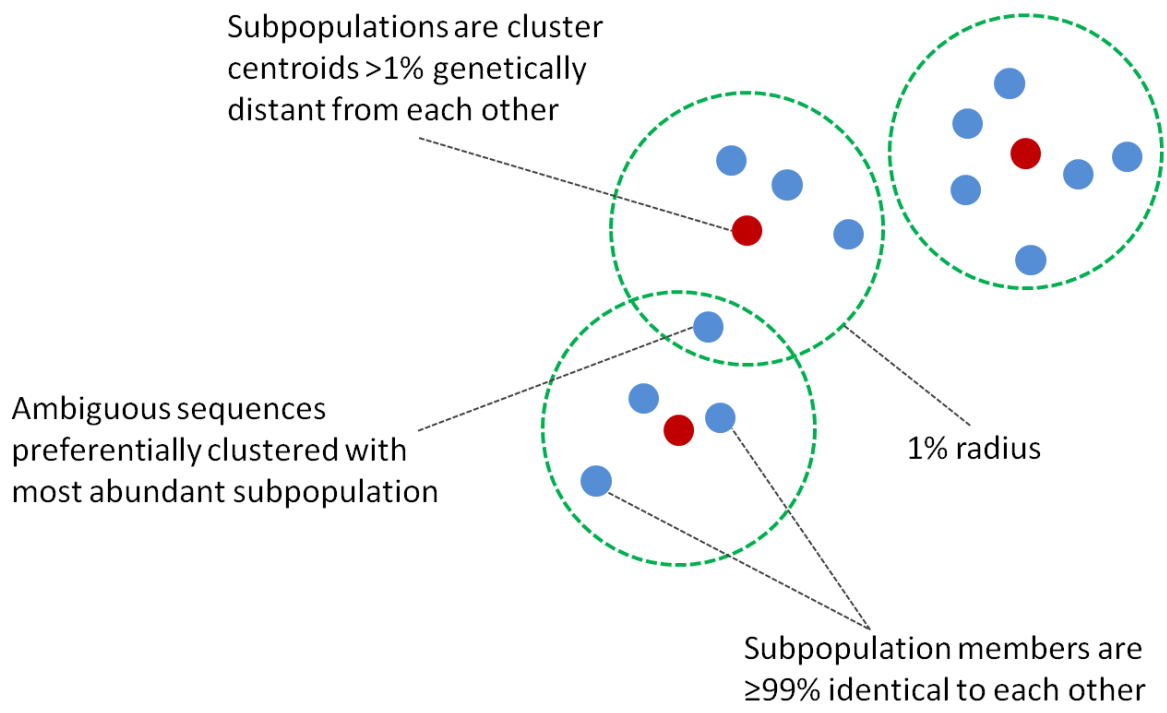


Figure 5.1. Schematic illustrating the principle of subpopulation identification using the UPARSE pipeline.

5.2.4 Estimation of patient infection date

A total of 500 randomly selected reads from the most abundant subpopulation in patient first time point samples only were submitted to the Poisson-Fitter tool in order to avoid over-estimating infection length because of multiple subpopulations (Giorgi, Funkhouser et al. 2010). If samples were deemed to have met the assumption of having a star-like phylogeny then the number of days to the most common recent ancestor (MRCA) was used as the date of infection. The median time to MRCA for those patients fitting a star-like phylogeny was used for samples where the assumption was not met.

5.2.5 Tropism

For each subpopulation identified in a sample, all original, ungapped unique reads within each subpopulation was submitted to the geno2pheno web tool designed for 454 reads (http://454.geno2pheno.org/index.php/) for tropism determination.

5.2.6 Transmitted drug resistance

For each sample subpopulation above the lower margin of error limit of 1%, the drug resistance profile of each unique read within that subpopulation was ascertained using the Stanford HIVdb tool (http://sierra2.stanford.edu/sierra/servlet/JSierra), with reference to the World Health Organization 2009 List of Mutations for Surveillance of Transmitted Drug Resistant HIV Strains (Bennett, Camacho et al. 2009). Due to the length of the *pol* amplicon, only resistance to NRTIs and NNRTIs was assessed.

5.2.7 Transmission chain samples

Pol Sanger consensus sequences for all 36 patients, together with 531 sequences obtained from the same London based cohort were subjected to phylogenetic analysis after stripping drug resistance associated codons based upon the Bennett 2009 list (Bennett, Camacho et al. 2009). Potential clusters were identified by constructing trees using PhyML3.0 (Guindon, Dufayard et al. 2010) with a General Time Reversible (GTR) model of substitution with proportion of invariable sites and gamma shape parameter estimated from the data, and then identifying clusters on the basis of having ≥95% bootstrap support from 1000 bootstrap replicate approximate likelihood trees constructed in FastTree2.1, and a mean pairwise Hamming distance ≤0.045. The same process was carried out on the *pol* and *env* deep-

sequencing amplicon subpopulations in order to identify clusters of non-primary subpopulations that may have been obscured in the Sanger *pol* sequence trees.

5.3 Results

5.3.1 Sample preparation and read generation

A total of 65 plasma samples were available for analysis (median viral load 5.2 log10 copies/ml, IQR: 4.8-5.7). A median of 5964 (IQR: 5402-6593) *pol* and 4650 (IQR: 4066-5135) *env* reads were available for each sample pre-quality filtering. After quality filtering using a maximum expected error threshold as implemented in UPARSE, a median of 3012 (IQR: 2527-3540) *pol* and 2888 (IQR: 2201-3411) *env* reads were available for analysis per sample.

5.3.2 Subpopulation identification

For the 36 patients with deep-sequencing results available there was a median of 1 subpopulation for *pol* (IQR: 1-2) and 2 subpopulation for *env* (IQR: 1-2) across all time points. The clone control had 1 subpopulation in *pol* and 1 subpopulation in *env* at the 1% read prevalence cut-off. The clone control had 3 subpopulations identified for *pol* at lower prevalence (0.5%, 0.4% and 0.2%) and 1 subpopulation identified for *env* at lower prevalence (0.9%), but these percentage abundances dropped further when using the lower bound of the margin of error (Lohr 2009). BLAST was used to identify potential sources of clone control contamination within the overall patient samples, and identical matches were found to other subpopulations within the same pool, suggesting that these reads were the result of crossover between samples, possibly due to MID tag errors (Carlsen, Aas et al. 2012). It should be emphasized that this was not the result of PCR contamination between samples, as the clone control was amplified in a separate reaction at a different time to the patient samples.

Putative subpopulations were subjected to the margin of error minimum bound filter and BLAST analysis to remove subpopulations present at low abundance or potentially present as a result of low-level MID tag contamination. This process removed 326/456 (71.5%) *pol* subpopulations across all samples (318 due to margin of error correction filtering, 8 due to BLAST filtering) present at a median abundance of 0.37% (IQR: 0.17-0.67%), and 385/624 (61.7%) *env* subpopulations across all samples (381 due to margin of error correction filtering, 4 due to BLAST filtering) present at a median abundance of 0.37% (IQR 0.16-0.76%). At the earliest sampled time point, median 139 days (IQR: 125.5-156 days) after estimated infection date as estimated using the Poisson-Fitter tool, 11/36 (30.6%) patients had evidence of >1 subpopulation in *pol* and 17/36 (47.2%) patients had evidence of >1 subpopulation in *env* (Table 1). For the 30/65 (53.8%) patient time points with >1 *pol* subpopulation, the average pairwise genetic distance between subpopulations was a median Hamming distance of 1.3% (IQR: 1.1-2.1%), and for the 41/65 (63.1%) patient time points with >1 *env* subpopulation, these subpopulations were a median Hamming distance of 1.4% (IQR: 1.2-2.3%) away from each other (measured using the 'centroid' seed sequence for each subpopulation) (N.B. by definition, the median distance between subpopulation must be >1%, since this is the distance specified in the UPARSE pipeline). For context, comparing patient 1 primary subpopulation in the first time point with the primary subpopulation in the fifth time point, over 57 weeks later, revealed a Hamming distance of 0.3% in *pol* and 1.3% in *env* respectively (N.B. the UPARSE parameters used to identify subpopulations specified that centroids must be >1% genetic distance from each other, but this does not apply to subpopulations from different time points).

Patient	Diagnosis date	CD4 count	Avidity index	Time point	Sample identifier	Sample date	Viral load	Pol		Env		Subpopulations	
								Pre-qc reads	Post-qc reads	Pre-qc reads	Post-qc reads	Pol	Env
Patient 1	10/06/2004	560	0.163	t0	Pool4_TC1	15/09/2004	36672	4722	2052 (43)	3781	2198 (58)	1	4
				t1	Pool4_TC2	17/12/2004	24844	4876	2124 (44)	3980	1992 (50)	2	9
				t2	Pool3_TC3	11/03/2005	39971	5808	2726 (47)	5759	3136 (54)	1	10
				t3	Pool3_TC6	15/06/2005	49809	5398	2387 (44)	5135	2759 (54)	2	7
				t4	Pool4_TC5	21/10/2005	35732	4954	2103 (42)	4425	2160 (49)	2	8
Patient 2	06/10/2004	302	0.234	t0	Pool3_TC1	30/11/2004	74842	6218	3068 (49)	5075	3042 (60)	1	1
				t1	Pool2_TC2	18/02/2005	51151	5074	2307 (45)	3389	1659 (49)	1	1
				t2	Pool3_TC4	04/05/2005	78929	6031	2800 (46)	5308	2710 (51)	2	2
				t3	Pool3_TC5	23/05/2005	90618	6190	2922 (47)	4837	2448 (51)	1	2
Patient 3	19/10/2004	359	0.161	t0	Pool2_TC1	19/10/2004	297620	6553	2999 (46)	3594	926 (26)	2	4
Patient 4	08/02/2005	663	0.144	t0	Pool4_TC3	01/03/2005	455309	4849	2838 (59)	4810	2476 (51)	2	2
				t1	Pool1_TC3	06/04/2005	506240	6798	4002 (59)	4503	1909 (42)	3	3
				t2	Pool1_TC4	27/04/2005	243047	5962	3584 (60)	4412	2056 (47)	4	3
				t3	Pool1_TC5	23/06/2005	380285	5694	3474 (61)	4019	1906 (47)	4	4
				t4	Pool3_TC7	22/08/2005	88451	6853	3982 (58)	4590	1850 (40)	4	6
				t5	Pool4_TC4	30/09/2005	21545	5438	3182 (59)	4654	2102 (45)	3	8
Patient 5	10/02/2005	348	0.328	t0	Pool1_TC1	10/02/2005	146868	6742	2860 (42)	4010	2763 (69)	1	1
Patient 6	16/02/2005	628	0.583	t0	Pool1_TC2	16/02/2005	281955	4747	2659 (56)	4227	2816 (67)	1	1
Patient 7	10/03/2005	355	0.238	t0	Pool5_TC1	16/03/2005	>1,000,000	7634	2613 (34)	3704	1665 (45)	2	2
Patient 8	17/05/2005	278	0.35	t0	Pool5_TC2	10/10/2005	150875	7643	3710 (49)	3646	2494 (68)	2	2
				t1	Pool4_TC6	14/11/2005	187633	5390	2569 (48)	5405	3382 (63)	2	3
				t2	Pool5_TC3	12/12/2005	626596	8414	4264 (51)	4293	2911 (68)	1	2
Patient 9	15/06/2005	634	0.17	t0	Pool3_TC8	19/10/2005	58433	5704	1912 (34)	5514	3214 (58)	2	13
				t1	Pool3_TC10	20/12/2005	37387	6081	1997 (33)	5962	3291 (55)	1	13
				t2	Pool4_TC7	22/03/2006	52619	5891	1241 (21)	5133	2949 (57)	2	12
				t3	Pool4_TC10	22/08/2006	49123	5047	1193 (24)	4751	2776 (58)	1	7
Patient 10	26/07/2005	283	0.52	t0	Pool2_TC3	18/08/2005	80164	6028	2364 (39)	4492	2432 (54)	1	1
Patient 11	28/12/2005	412	0.33	t0	Pool2_TC4	28/12/2005	1391869	5835	2803 (48)	4904	2919 (60)	1	1
Patient 12	06/01/2006	587	0.24	t0	Pool5_TC4	06/01/2006	>10,000,000	8801	4783 (54)	4727	1894 (40)	1	1
Patient 13	17/01/2006	1066	0.47	t0	Pool1_TC6	17/01/2006	4452318	5717	2525 (44)	4893	3212 (66)	1	4
				t1	Pool1_TC7	27/01/2006	1878960	6679	2816 (42)	5539	3480 (63)	4	5
				t2	Pool1_TC8	10/02/2006	1390722	5413	2282 (42)	4660	2966 (64)	4	5
				t3	Pool3_TC11	30/03/2006	155060	6593	2757 (42)	5842	3370 (58)	4	6
				t4	Pool3_TC13	25/04/2006	130031	6487	2798 (43)	5699	3165 (56)	4	13
				t5	Pool1_TC10	10/05/2006	519051	6412	2732 (43)	5807	3525 (61)	4	12
Patient 14	04/04/2006	295	0.7	t0	Pool2_TC5	03/04/2006	261986	4783	1682 (35)	3973	1739 (44)	1	1
Patient 15	22/08/2006	1057	0.4	t0	Pool2_TC6	21/08/2006	50850	5422	1942 (36)	4013	2167 (54)	1	1
Patient 16	29/09/2006	378	0.6	t0	Pool4_TC11	28/09/2006	327130	6038	3807 (63)	5066	2165 (43)	8	1
Patient 17	09/11/2006	502	0.6	t0	Pool4_TC13	09/11/2006	696163	5804	2085 (36)	4650	2724 (59)	1	1
Patient 18	20/03/2007	1036	0.5	t1	Pool1_TC11	15/03/2007	1652460	6518	3630 (56)	5617	3800 (68)	1	1
Patient 19	10/04/2007	787	0.4	t0	Pool5_TC5	13/04/2007	107360	8509	3087 (36)	4224	1985 (47)	1	1
Patient 20	11/05/2007	1001	0.6	t0	Pool2_TC7	11/05/2007	268640	5903	2727 (46)	4779	2763 (58)	1	1
Patient 21	25/05/2007	764	0.6	t0	Pool2_TC8	25/05/2007	141578	4684	2469 (53)	3357	1620 (48)	1	4
Patient 22	31/05/2007	608	0.3	t0	Pool5_TC6	31/05/2007	>10,000,000	7858	4581 (58)	4243	2780 (66)	1	2
Patient 23	12/06/2007	399	0.4	t0	Pool4_TC14	12/06/2007	33082	5169	2253 (44)	3875	2173 (56)	1	2
Patient 23				t1	Pool3_TC14	15/06/2007	52299	4863	2485 (51)	5170	3197 (62)	1	1
Patient 23				t2	Pool3_TC15	21/08/2007	90528	5964	2828 (47)	5347	3508 (66)	1	1
Patient 24	28/06/2007	462	0.5	t0	Pool2_TC11	29/06/2007	130267	4236	1547 (37)	3734	640 (17)	4	3
Patient 25	01/11/2007	951	0.7	t0	Pool2_TC14	01/11/2007	299850	5923	2816 (48)	4495	1842 (41)	1	11
Patient 26	07/11/2007	272	0.7	t0	Pool5_TC10	12/11/2007	604026	8506	4696 (55)	5195	3232 (62)	2	5
Patient 27	23/04/2008	516	0.5	t0	Pool2_TC15	23/04/2008	94013	5246	2534 (48)	3651	2109 (58)	1	1
Patient 28	13/06/2008	339	0.5	t0	Pool5_TC13	13/06/2008	108093	8324	3641 (44)	3286	1078 (33)	1	2
Patient 29	30/07/2008	919	0.1	t0	Pool4_TC16	29/07/2008	40064	5862	3371 (58)	5202	3202 (62)	1	2
Patient 30	18/08/2008	494	0.3	t0	Pool5_TC14	18/08/2008	354973	8111	3659 (45)	3687	1951 (53)	1	1
Patient 31	17/09/2008	326	0.4	t0	Pool2_TC16	17/09/2008	8232216	5131	1685 (33)	4526	1563 (35)	2	1
Patient 32	17/10/2008	349	0.5	t0	Pool2_TC18	17/10/2008	4304339	6035	1385 (23)	4015	1027 (26)	3	2
Patient 33	21/01/2009	744	0.3	t0	Pool5_TC15	16/01/2009	40100735	6155	2647 (43)	4206	2774 (66)	1	1
Patient 34	28/01/2009	221	0.2	t0	Pool5_TC16	30/01/2009	>10,000,000	7726	4595 (59)	4207	2077 (49)	3	2
Patient 35	21/05/2009	806	0.3	t0	Pool4_TC17	21/05/2009	140230	5185	2704 (52)	5002	2227 (45)	1	1
Patient 35				t1	Pool3_TC16	29/05/2009	96840	6097	3062 (50)	4644	2228 (48)	1	1
Patient 35				t2	Pool4_TC18	12/06/2009	40174	5835	3032 (52)	5111	2379 (47)	1	1
Patient 36	19/06/2009	760	0.1	t0	Pool1_TC17	26/06/2009	872971	6911	2902 (42)	5177	3287 (63)	1	1
Patient 36				t1	Pool3_TC17	10/08/2009	61034	6756	2637 (39)	5233	3355 (64)	2	2
Patient 36				t2	Pool5_TC18	11/12/2009	87052	6124	2464 (40)	4224	2743 (65)	5	4
Patient 36				t3	Pool1_TC18	04/03/2010	182705	6093	2275 (37)	4867	3185 (65)	5	2

Table 5.1. Clinical data from diagnosis for 36 patients eligible for deep-sequencing. Avidity index is the guanidine based avidity assay result used to estimate length of HIV-1 infection. T0 sample is the first time point sample that deep-sequencing was carried out on, which was not always the same as the diagnosis sample. Subpopulations at T0 is

the number of subpopulations > 1% abundance in the first time point sample. Pre-qc read number refers to number of reads after mapping reads to *pol* or *env* regions based upon primer sequence. Post-qc reads refers to read number after imposing a read quality filtering step through the UPARSE pipeline, together with an open-reading frame minimum length filter to remove spurious reads generated through indels.

5.3.3 Patients with multiple time points

Pol: 6/9 patients with multiple time points available had single *pol* subpopulations at first time point, 3/9 had evidence of >1 *pol* subpopulation. Of those patients with single *pol* subpopulations at first time point, 4/6 went on to increase in subpopulation number in later time points, the remaining two maintained a single subpopulation across all time points, for follow-up periods of 10 and 3 weeks respectively.

Overall pairwise genetic diversity (i.e. combined across subpopulations) in *pol* increased from first to last time point for 7/9 (77.8%) patients, but the increase was not consistent between each time point for all patients (Fig 5.2).

Env: 3/9 patients with multiple time points had single *env* subpopulations at first time point, 6/9 had evidence of >1 *env* subpopulation. Of those patients with single *env* subpopulations at first time point, 2/3 went on to increase in subpopulation number in later time points, the infection of the remaining patient maintained a single subpopulation across all time points, for 3 weeks.

Overall pairwise genetic diversity (i.e. combined across subpopulations) in *env* increased from first to last time point for 6/9 (66.7%) patients, but the increase was not consistent between each time point for all patients (Fig 5.2).

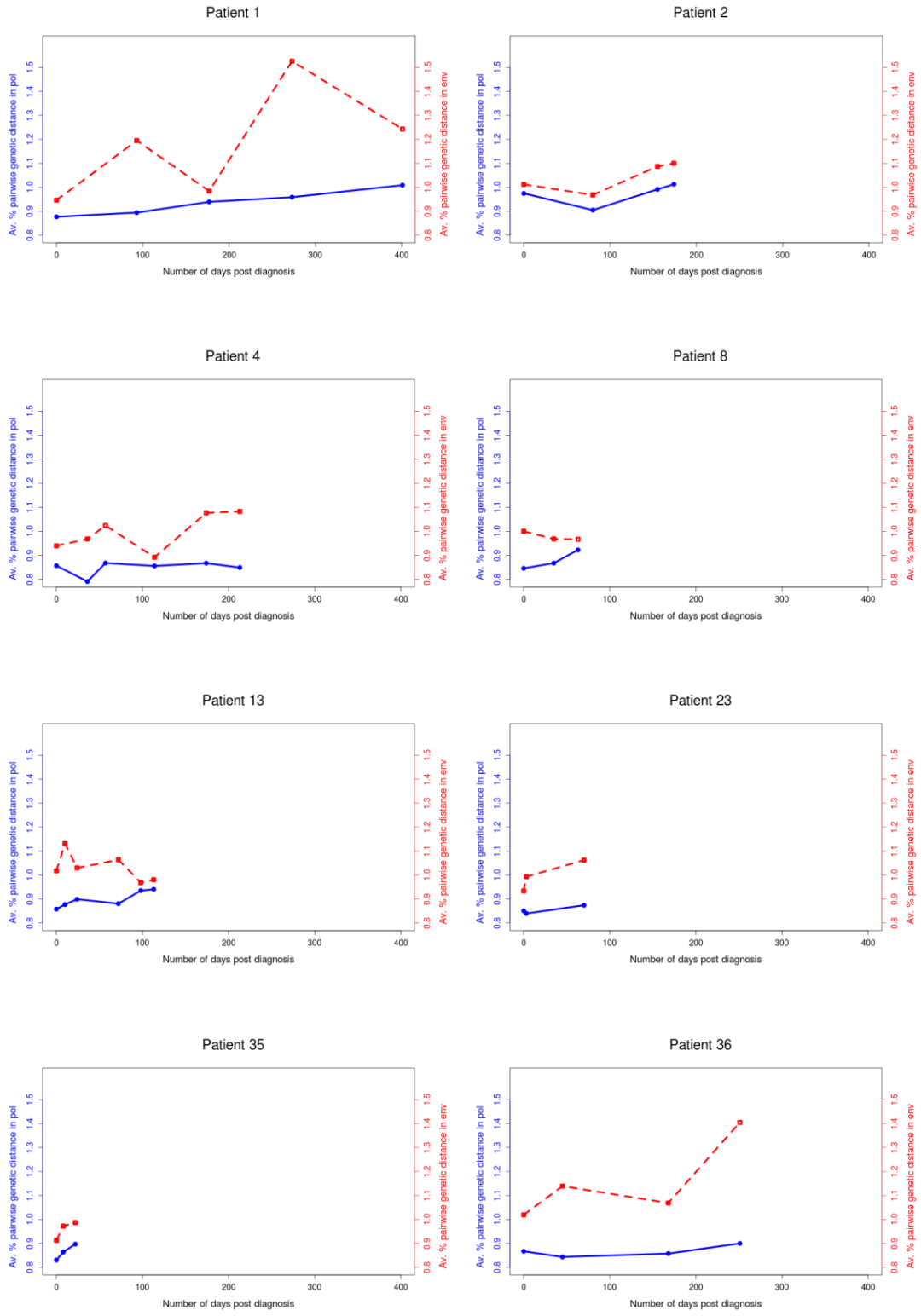
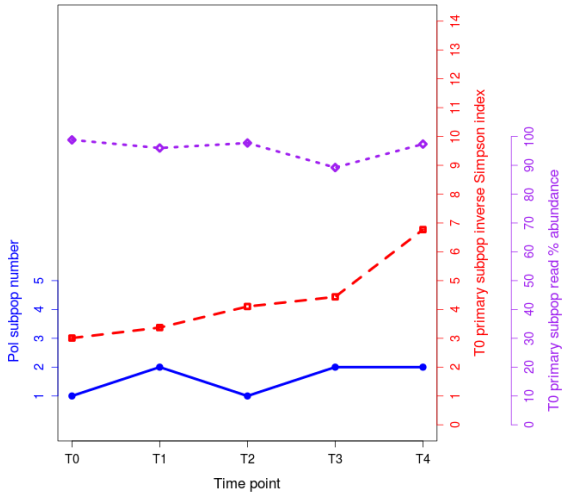


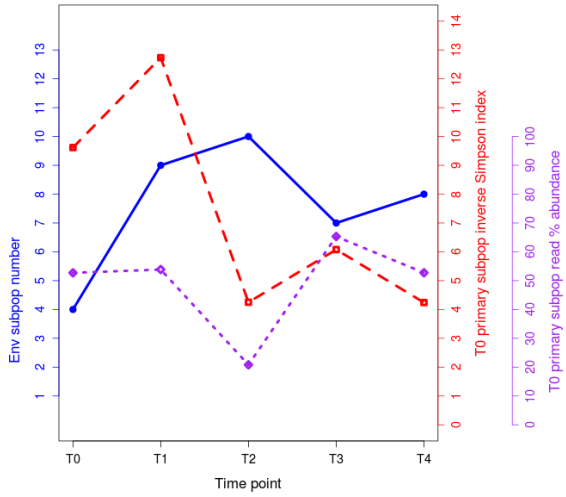
Figure 5.2. Plots showing overall changes in pairwise Hamming distance between reads across all subpopulations combined for the 9 patients with multiple time points available (blue solid lines for *pol*, and red dashed lines for *env*).

For each of the 9 patients with multiple time points, the inverse Simpson index of diversity was calculated for the primary subpopulation in the first time point and the most abundant closest matching subpopulation in subsequent time points using the Vegan package in R (Oksanen, Blanchet et al. 2013), where an increase in the index indicates an increase in subpopulation diversity (i.e. a greater number of unique reads, and fewer identical reads) (Fig 2.). Where the percentage read abundance of the primary subpopulation remained close to 100%, the diversity index and overall subpopulation number remained relatively stable. It is in patients 1, 4, 13 and 36 that display the most radical disruption in primary subpopulation read abundance and diversity within that read. In these four patients, primary subpopulation read abundance seems to inversely correlate with overall subpopulation number.

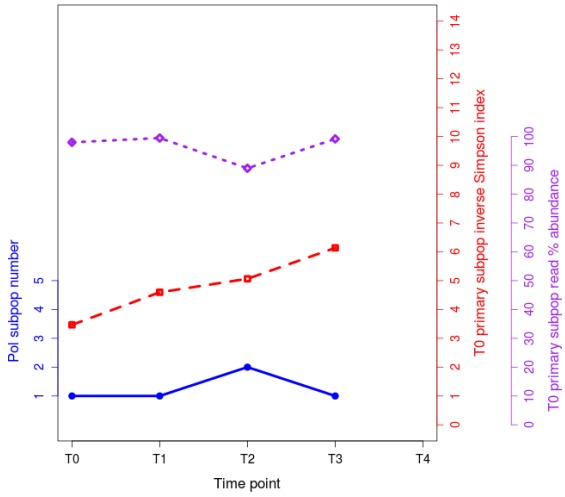
Figure 5.3. Plots showing changes in the total subpop(ulation) number in each patient time point, together with the inverse Simpson index for the primary subpopulation in t0 tracked through all time points (with higher values indicating greater diversity), and the % abundance of that primary subpopulation in each time point. For each patient, left panel shows results for *pol*, and right panel results for *env*. Axis colours are matched to plot colours for clarity.

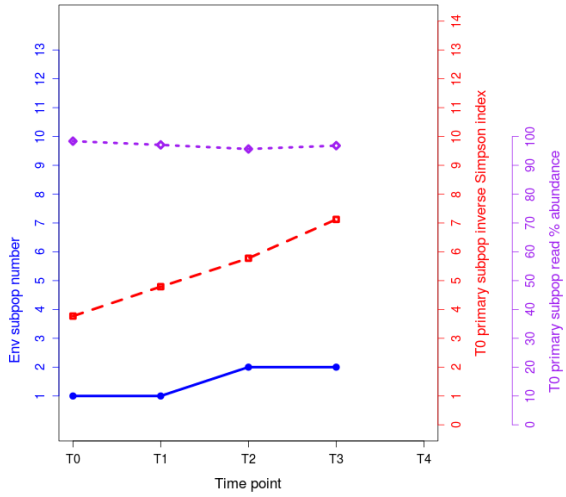
Patient 1



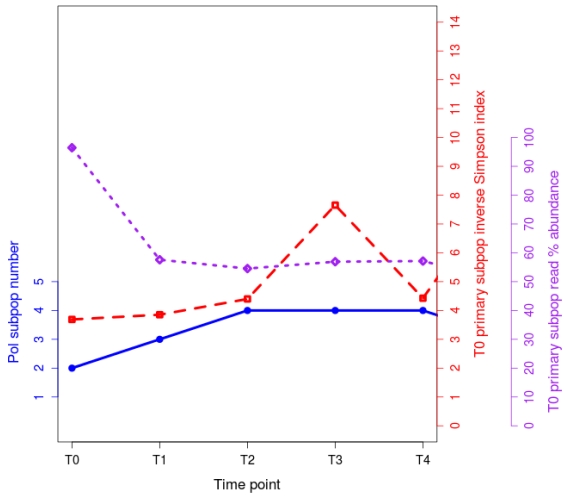


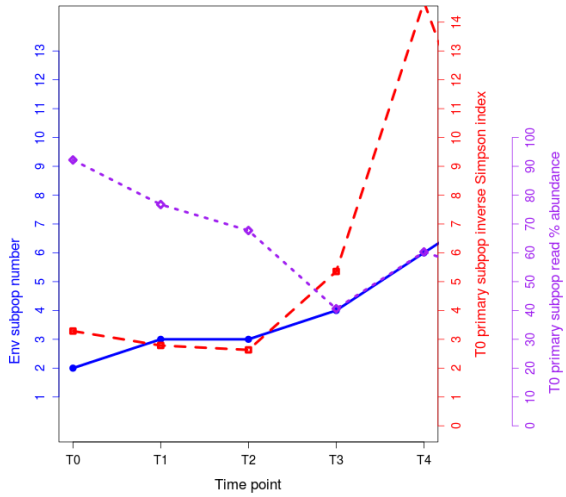
Patient 2



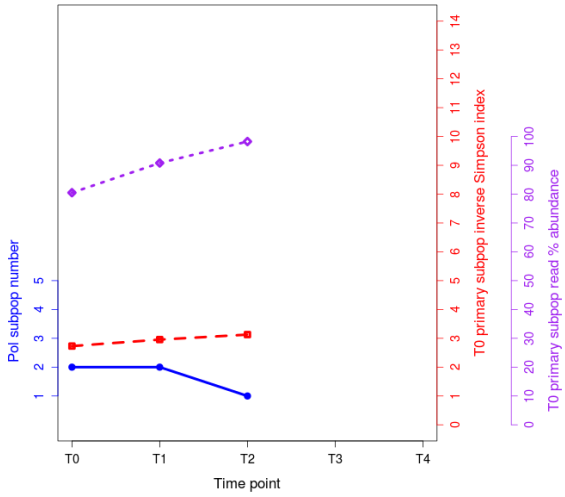


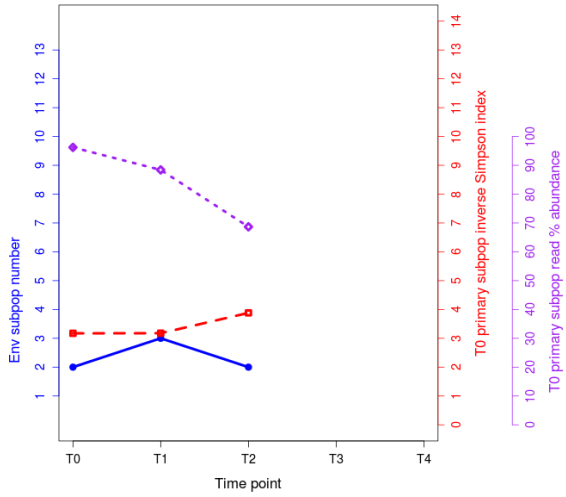
Patient 4



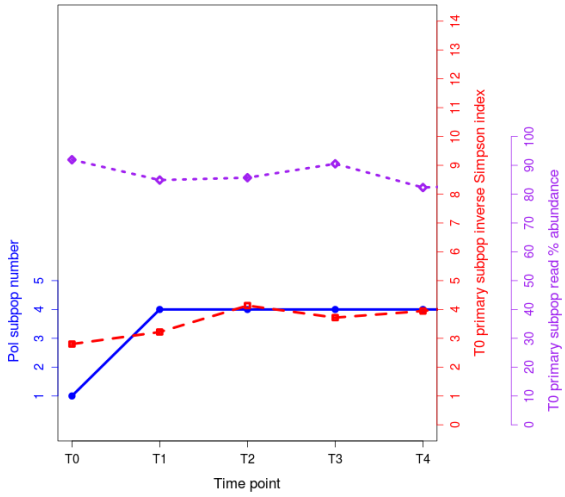


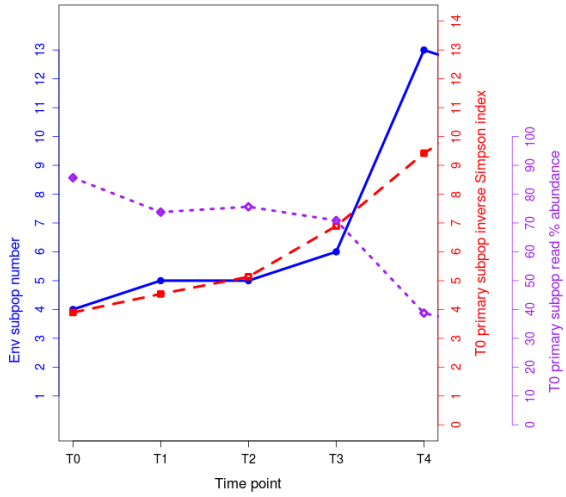
Patient 8



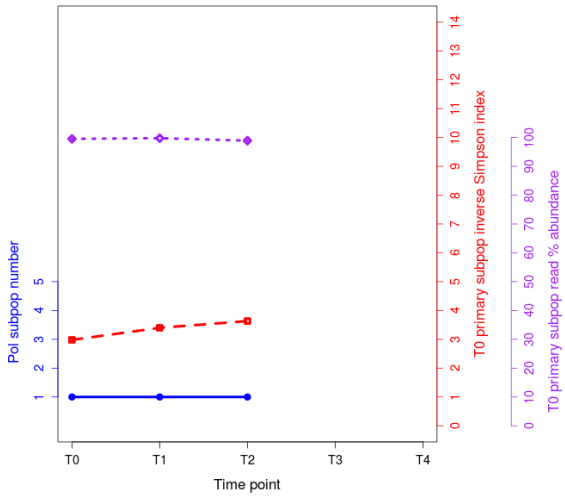


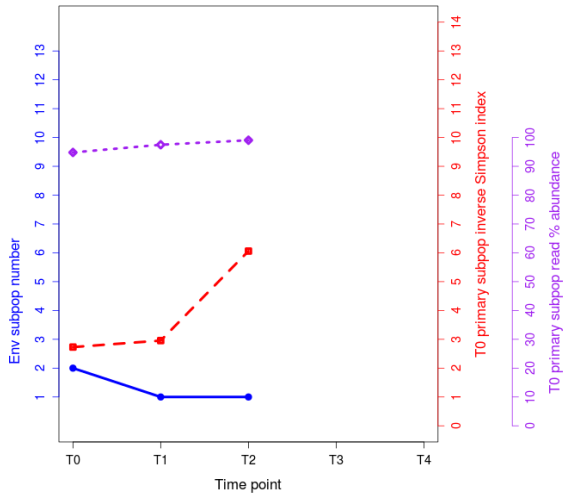
Patient 13

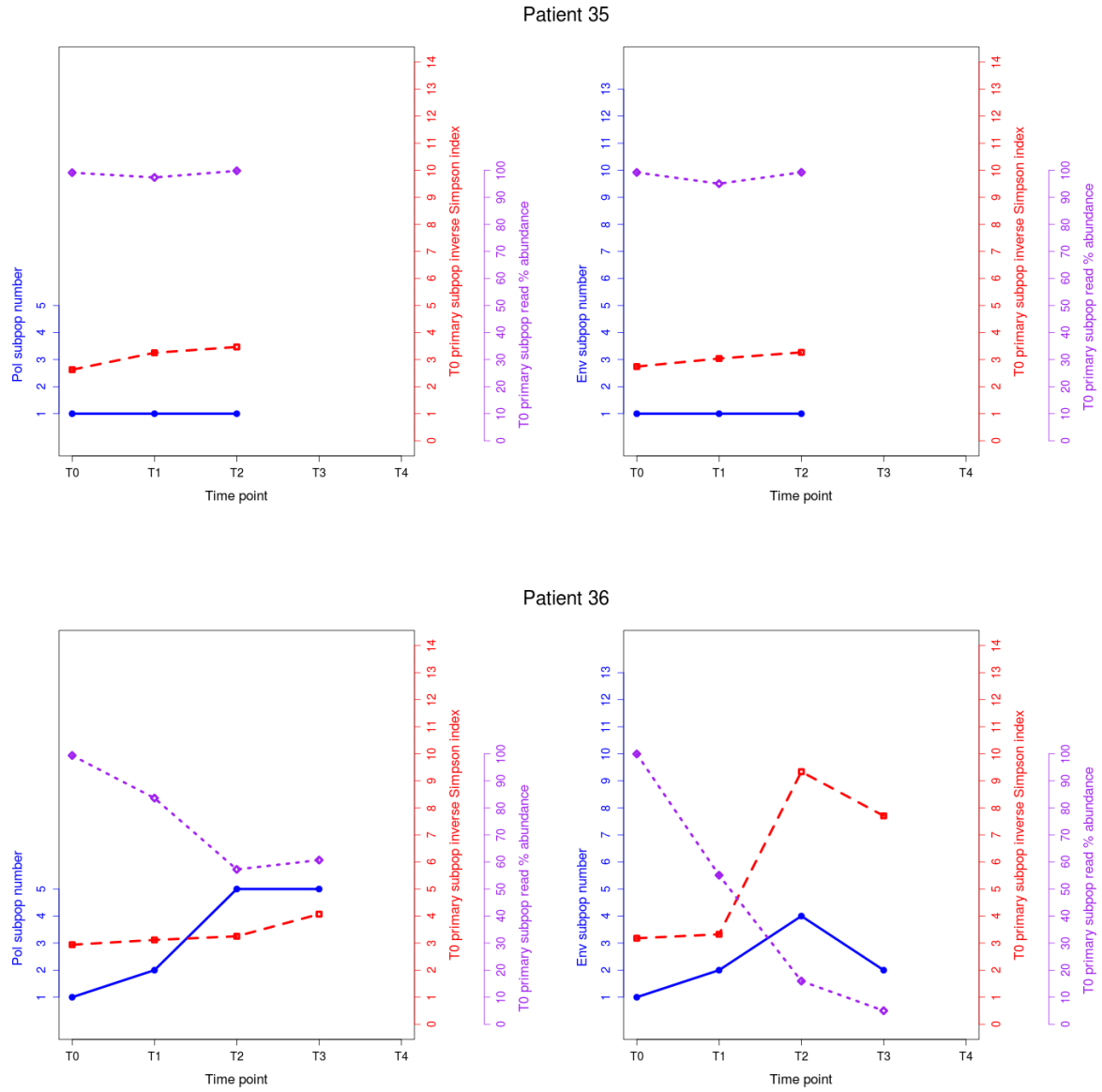




Patient 23







Rate of evolution of subpopulations: the rate of evolution seen in the primary *pol* and *env* subpopulations between first and last time point in the 9 patients with multiple samples, was a median 2.4 x 10^{-3} nucleotide substitutions per site per year (IQR: 0 to 7.6 x 10^{-3}) and 2.9 x 10^{-2} nucleotide substitutions per site per year (IQR: 1.3 x 10^{-2} to 4.6 x 10^{-2}) in *pol* and *env* respectively, based upon Hamming distance.

Subpopulation subtypes: The HIV-1 subtype did not agree between the predominant *pol* and *env* amplicons in 3/36 patient first time points. In 4/21 patients with >1 *pol* or *env*

subpopulation at their first or subsequent time points, the additional subpopulations were of different subtype to the primary subpopulation (either in *pol* or *env*).

Two of these patients displayed outgrowth of secondary/tertiary subpopulations of different subtype to the primary subpopulation over multiple time points:

- Patient 8 had a primary subtype B subpopulation in *env* in the first time point, with a secondary subtype A1 subpopulation present at 2.5% abundance. In the second time point, this A1 subpopulation had diversified into two subpopulations present at 6.9% and 3.9%, and by the third time point had merged into one subpopulation present at 30.8% abundance. The primary *pol* subpopulation remained subtype B across all three time points (Table S1.).

- Patient 36 had a primary subtype B subpopulation in both *pol* and *env* in the first time point (with a suggestion by the COMET subtyping algorithm to check for presence of CRF29_BF in the *pol* amplicon), with no evidence of secondary subpopulations above 1% abundance. By the second time point, *pol* had developed a secondary subtype B subpopulation at 11.1% (with a suggestion by the COMET subtyping algorithm to check for presence of CRF12_BF), and *env* had developed a secondary subpopulation at 44.2% abundance with an A1 subtype. By the third time point, *pol* had developed 5 subtype B subpopulations (at 57.3%, 15.6%, 3.4% and 3.2% abundance respectively), and a subtype D subpopulation at 3.2% (though subtype B and D are genetically highly similar, and so this subpopulation may also be subtype B, but with insufficient sequence to allow correct allocation of subtype (Robertson, Anderson et al. 1999)), and the *env* A1 subpopulation had grown to an abundance of 75.3%, with 3 secondary subpopulations: 2 subtype F1 subpopulations, present at 16.0% and 5.1% abundance, and another subtype A1 subpopulation present at 2.5%. In time point 4, *pol* was found

to have 5 subtype subpopulations in *pol* at 60.1%, 19.5%, 3.0%, 2.1% and 2.0%, whereas the subtype A1 subpopulation in *env* was found to be present at 94.1% abundance and just one secondary subpopulation, subtype B, at 5.0% (Appendix 3).

5.3.4 Genetic distance sensitivity analysis

The genetic distance used to separate initial subpopulation 'seeds' (centroids) and subsequently cluster highly related reads within a sample was altered to assess the effect on the overall subpopulation number identified within each sample. An original >1% genetic distance between putative subpopulations, and a corresponding clustering of all sequences with ≥99% genetic identity to the initial subpopulation 'seed' sequence, was replaced by a >3% genetic distance cut-off for initial subpopulation 'seeds', combined with a corresponding ≥97% genetic identity cut-off. Results indicated a reduced number of subpopulations for both *pol* and *env* identified at the more generous cut-off (Table 5.4).

			1% subpopulations		3% subpopulations	
Patient	**Time point**	**Sample identifier**	**Pol**	**Env**	**Pol**	**Env**
Patient 1	t0	Pool4_TC1	1	4	1	1
	t1	Pool4_TC2	2	9	1	1
	t2	Pool3_TC3	1	10	1	1
	t3	Pool3_TC6	2	7	1	1
	t4	Pool4_TC5	2	8	1	1
Patient 2	t0	Pool3_TC1	1	1	1	1
	t1	Pool2_TC2	1	1	1	1
	t2	Pool3_TC4	2	2	1	1
	t3	Pool3_TC5	1	2	1	1
Patient 3	t0	Pool2_TC1	2	4	1	2
Patient 4	t0	Pool4_TC3	2	2	1	1
	t1	Pool1_TC3	3	3	1	1
	t2	Pool1_TC4	4	3	1	1
	t3	Pool1_TC5	4	4	1	1
	t4	Pool3_TC7	4	6	1	1
	t5	Pool4_TC4	3	8	1	1
Patient 5	t0	Pool1_TC1	1	1	1	1
Patient 6	t0	Pool1_TC2	1	1	1	1
Patient 7	t0	Pool5_TC1	2	2	2	1
Patient 8	t0	Pool5_TC2	2	2	2	2
	t1	Pool4_TC6	2	3	2	2
	t2	Pool5_TC3	1	2	1	2
Patient 9	t0	Pool3_TC8	2	13	1	2
	t1	Pool3_TC10	1	13	1	1
	t2	Pool4_TC7	2	12	1	1
	t3	Pool4_TC10	1	7	1	3
Patient 10	t0	Pool2_TC3	1	1	1	1
Patient 11	t0	Pool2_TC4	1	1	1	1
Patient 12	t0	Pool5_TC4	1	1	1	1
Patient 13	t0	Pool1_TC6	1	4	1	2
	t1	Pool1_TC7	4	5	2	2
	t2	Pool1_TC8	4	5	2	2
	t3	Pool3_TC11	4	6	2	2
	t4	Pool3_TC13	4	13	2	2
	t5	Pool1_TC10	4	12	2	1
Patient 14	t0	Pool2_TC5	1	1	1	1
Patient 15	t0	Pool2_TC6	1	1	1	1
Patient 16	t0	Pool4_TC11	8	1	1	1
Patient 17	t0	Pool4_TC13	1	1	1	1
Patient 18	t1	Pool1_TC11	1	1	1	1
Patient 19	t0	Pool5_TC5	1	1	1	1
Patient 20	t0	Pool2_TC7	1	1	1	1
Patient 21	t0	Pool2_TC8	1	4	1	1
Patient 22	t0	Pool5_TC6	1	2	1	1
Patient 23	t0	Pool4_TC14	1	2	1	1
Patient 23	t1	Pool3_TC14	1	1	1	1
Patient 23	t2	Pool3_TC15	1	1	1	1
Patient 24	t0	Pool2_TC11	4	3	1	1
Patient 25	t0	Pool2_TC14	1	11	1	1
Patient 26	t0	Pool5_TC10	2	5	2	2
Patient 27	t0	Pool2_TC15	1	1	1	1
Patient 28	t0	Pool5_TC13	1	2	1	2
Patient 29	t0	Pool4_TC16	1	2	1	1
Patient 30	t0	Pool5_TC14	1	1	1	1
Patient 31	t0	Pool2_TC16	2	1	1	1
Patient 32	t0	Pool2_TC18	3	2	2	2
Patient 33	t0	Pool5_TC15	1	1	1	1
Patient 34	t0	Pool5_TC16	3	2	1	2
Patient 35	t0	Pool4_TC17	1	1	1	1
Patient 35	t1	Pool3_TC16	1	1	1	1
Patient 35	t2	Pool4_TC18	1	1	1	1
Patient 36	t0	Pool1_TC17	1	1	1	1
Patient 36	t1	Pool3_TC17	2	2	2	2
Patient 36	t2	Pool5_TC18	5	4	3	2
Patient 36	t3	Pool1_TC18	5	2	3	2

Table 5.4. Comparing subpopulation numbers for *pol* and *env* using a 1% and 3% genetic distance cut-off in the UPARSE pipeline.

5.3.5 Tropism

Overall 19/238 (8.0%) of the stringently filtered *env* subpopulations submitted to the geno2pheno 454 web tool was found to have >0% of predicted percentage of X4-tropic virus at an FPR of 5%, ranging from 1.4% to 8.2%, with one subpopulation having a percentage of 61.9%. This subpopulation belonged to the first time point of patient 35 (single *env* subpopulations across all three time points), highly likely to have been sampled from Fiebig stage II (patient diagnosed with p24 Ag^{+}/Ab^{-} results) (Cohen, Gay et al. 2010). By the second time point 8 days later, a single nucleotide mutation in the subpopulation representative sequence resulted in an amino acid change from glutamic acid to glycine at position 24 of the V3 loop (20-WYTP*E*QITGDI-30 to 20-WYTP*G*QITGDI-30), changing the of predicted percentage of X4-tropic virus at an FPR of 5% from 61.9% to 0.3%.

5.3.6 Transmitted drug resistance

Resistance to NRTIs and NNRTIs was assessed for all 129 *pol* subpopulations containing >1% of reads in each sample (margin of error corrected), using the Sierra web service available on the Stanford HIV Drug Resistance Database website. Overall, 101/129 subpopulations had drug resistance, with a median of 3 drug resistance mutations (IQR: 1-5). Of these 101 drug resistance subpopulations, there was a median of 2.8 reads per drug resistance mutation (IQR: 1-5.6). Due to the error prone process of deep-sequencing, it is

possible that some of these mutations were introduced as part of the experimental procedure, and indeed, the average reads per drug resistance mutation appeared to reveal a division between those subpopulations with <10 reads per drug resistance mutation (116 subpopulations, median 2.0 reads per mutation, IQR: 0.8 to 4.6 reads per mutation) and those with >150 reads per mutation (13 subpopulations, median 931.5 reads per mutation, IQR: 258.2 to 1419.5 reads per mutation), possibly suggesting the former represent mutations that are either present at extremely low levels *in the majority of treatment naive individuals* in the cohort, or that the mutations are spurious artefacts of the 454 sequencing process.

NRTI resistance: Taking forward only the 13 subpopulations with >150 average number of reads per mutation, 7/13 subpopulations had a T215S mutation, 5 from one patient with 4 time points available across 25 weeks of follow-up (with the third time point have two subpopulations, both carrying the T215S mutation), and the remaining two from patients with only one time point available. 1/13 subpopulations had K219E, 3/13 had K219N (subpopulations from the same patient time point), and 1/13 subpopulations had a K219Q mutation.

NNRTI resistance: 1/13 subpopulations had a Y188C mutation.

5.3.7 Transmission chain samples

A phylogenetic tree was constructed with PhyML3.0 (Guindon, Dufayard et al. 2010) using the 36 patient *pol* Sanger sequences, together with 531 sequences obtained from the same London based cohort. All sequences had drug resistance associated codons removed prior to analysis (Bennett, Camacho et al. 2009). Clusters were identified on the basis of having ≥95% bootstrap support from 1000 bootstrap replicate approximate likelihood trees constructed in FastTree2.1, and a mean pairwise Hamming distance ≤0.045. Two potential

clusters involving patients from the deep-sequencing cohort were identified, one involving 3 patients located within a subtype B cluster of 12 sequences with 100% bootstrap support and mean pairwise Hamming distance of 0.009, and one in a CRF02_AG cluster of 2 sequences with 100% bootstrap support and mean pairwise Hamming distance of 0.017. Further phylogenetic trees were constructed using the *pol* or *env* subpopulations from the 36 deep-sequencing patients only.

Subtype B cluster: 3 patients involved in a subtype B cluster of 12 individuals in the Sanger *pol* sequence tree had deep-sequencing results available: patient 8, patient 21 and patient 22.

Patient 8 showed 2 *pol* subpopulations at t0 (80.5% and 16.7% abundance respectively, both subtype B) and t1 (90.8% and 2.4% abundance respectively, both subtype B), but only one *pol* subpopulation in t2.

Patient 21 showed a single *pol* subpopulation at t0 (subtype B).

Patient 22 showed a single *pol* subpopulation at t0 (subtype B).

In the phylogenetic tree based upon deep-sequencing subpopulations, the secondary *pol* subpopulations detected in patient 8 at t0 and t1 were found to cluster with the *pol* subpopulation of patients 21 and 22 (bootstrap support score 97%, mean pairwise Hamming distance 0.004, Fig 4).

Of note, the first time point sampled for deep-sequencing for patient 8 was obtained 21 weeks after diagnosis - comparison of the first time point with the Sanger *pol* sequence generated at diagnosis indicates the primary subpopulation sampled at diagnosis had become the secondary subpopulation by the first time point submitted to deep-sequencing (21 weeks later).

In the *env* phylogenetic tree constructed from deep-sequencing subpopulations, the secondary *env* subpopulations detected in patient 8 at t0, t1 and t2 clustered with the *env* subpopulations detected in patients 21 and 22 (bootstrap support score 95% and the mean pairwise Hamming distance 0.034). In patient 8, in contrast to the *pol* secondary subpopulation, the *env* secondary subpopulation showed a steady increase over subsequent time points, suggesting it may have eventually emerged as the primary *env* subpopulation (Fig 5.)

A slightly larger cluster included four *env* subpopulations from patient 36: the subtype A1 secondary subpopulation detected at t1 (which became the primary subpopulation in t2 and t3), and the subtype A1 primary subpopulations from t2 and t3 (Fig 5). The cluster showed 100% bootstrap support with a mean pairwise Hamming distance of 0.053.

CRF02_AG cluster: 2 patients involved in a CRF02_AG transmission pair in the Sanger *pol* sequence tree had deep-sequencing results available: patient 26 and patient 29.

Patient 26: showed 2 *pol* subpopulations at t0 (both CRF02_AG).

Patient 29: showed one *pol* subpopulation at t0 (CRF02_AG).

Subpopulations were found to cluster in both the *pol* and *env* trees, this was with a bootstrap support score of 81% and 100% and a mean pairwise Hamming distance of 0.027 and 0.021 in *pol* and *env* respectively (Figs 5.5a and 5.5b).

Fig 5.4a.PhyML tree for *pol* subpopulations across all time points for all 36 patients. Clusters identified in Sanger sequence *pol* trees are highlighted.

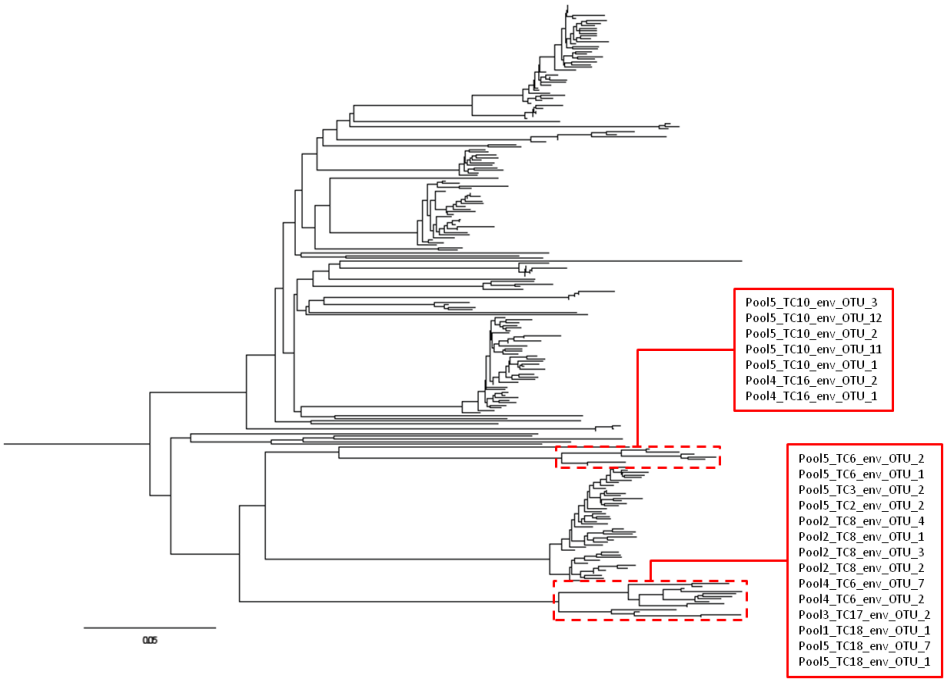


Fig 5.5b. PhyML tree for *env* subpopulations across all time points for all 36 patients. Clusters identified in Sanger sequence *pol* trees are highlighted.

5.4 Discussion

This study set out to examine the intra-host viral complexity in patients with early HIV-1 infection, using a deep-sequencing approach to delve into the quasi-species dynamics taking place at a level beyond the reach of traditional Sanger sequencing. Analysis of the first time point results across the 36 patients indicated that 11/36 (30.6%) and 17/36 (47.2%) individuals had >1 subpopulation in *pol* and *env* respectively. Given that the time of infection is unknown for these patients, but is estimated to be 139 days on average using the Poisson-Fitter tool, and it is clear that subpopulation numbers (as defined as >1% genetically distant from one another) change between time points that are substantially less than 139 days, it is difficult to assess the extent to which these subpopulations are evolving from one founding virus, or >1 founding virus. The 1/4 and 2/8 instances of secondary subpopulations of different subtype to the primary subpopulation in the first time point for *pol* and *env* respectively, seem to strongly indicate that the secondary subpopulations did not evolve from the primary subpopulations in each case, but was more likely the result of dual/superinfection.

In terms of subpopulation dynamics over time, this study identified a mixture of individuals exhibiting uniform subpopulation number over time, and individuals where substantial growth of initially minor subpopulations occurred over the time points. Perhaps most interesting were the two subjects (patients 8 and 36) that displayed substantial growth of a subpopulation that was in the minority, or not even present at ≥1% abundance, in the first time point, but then subsequently showed a large increase in abundance. Furthermore, these subpopulation outgrowths occurred in *env* but not to the same extent in *pol*, suggestive of recombination occurring over time somewhere between the two amplicon regions. This recombination and potential change in fitness could potentially be further investigated using transmissibility and replicative fitness assays (Njai, Gali et al. 2006), together with whole

genome sequencing of virus sampled from patients 8 and 36 to determine the existence of breakpoint locations. The link between changes in subpopulation number and changes in diversity over time points was not uniform cut across the 9 patients. In some instances (patients 1, 4 , 13 and 36) increases in diversity within the primary *env* subpopulation seemed to occur in tandem with increases in subpopulation number, possibly reflecting host pressure on the primary subpopulation, forcing it to explore multiple evolutionary pathways, with successful 'child' subpopulations developing into subpopulations >1% genetically distant from the 'parent' subpopulation; corresponding drops in the time point percentage abundance for the t0 primary subpopulations fit with this interpretation. Other patients appear to have little evolution in either *pol* or *env*, and it is interesting to speculate on the differences between these patient subsets, i.e. potential effects of Human Leukocyte Antigen (HLA) types of the patients.

It is also intriguing that the subtype A1 *env* subpopulations that seemed to be responsible for the substantial growths in patients 8 and 36 were highly related, and related to viruses in two other individuals that displayed strong indications of being epidemiologically linked to patient 8 on the basis of *pol* Sanger consensus sequencing (patients 21 and 22). One possible scenario could be that patient 8 was infected by an individual with a dual subtype B and subtype A1 infection, but that the A1 subtype virus was initially restricted or out-competed by the subtype B virus. Then, over time, possibly through recombination, the viral quasi-species developed into a primary subpopulation with a subtype B *pol* region and subtype A1 *env* region (the subtype of other regions remain to be elucidated). This virus was then transmitted to patients 21 and 22 (whether directly or not is impossible to determine, as there were 12 individuals in the *pol* Sanger consensus sequence cluster, and there may have been further individuals involved in the cluster who have not been diagnosed by the clinic in question). The virus was transmitted to patient 36 at a later time point, but as with patient 8, it

may have been initially restricted or outcompeted in some way by a co-founder virus but went on to outgrow that initial virus rapidly to become the primary subpopulation 24 weeks later, possibly with partial recombination once again, as the initial primary subtype B subpopulation in *pol* remains the primary subpopulation through all four time points (the likelihood of there being a super-infection event between t0 and t1 for patient 36 seems to be slim given the 6 week gap between samples, but cannot be ruled out).

In terms of one potential explanation of why particular viruses may play a more prominent role at the transmission founder stage and then go on to be outcompeted or evolve away from their initial form is the tropism of the virion, i.e. which co-receptor is preferred for host cell entry. Viruses that preferentially bind the CCR5 receptor appear to dominate early in HIV infection (Van't Wout, Kootstra et al. 1994; Salazar-Gonzalez, Salazar et al. 2009), with CXCR4 viruses becoming more prevalent later in infection (Schuitemaker, Koot et al. 1992; Connor, Sheridan et al. 1997), though it is still unclear if this is a cause of or correlation with disease progression. In previous studies, CXCR4 tropic viruses have been found at low levels in individuals prior to treatment with the co-receptor antagonist maraviroc (Archer, Rambaut et al. 2010; Bunnik, Swenson et al. 2011). It is debated to what extent these low-frequency CXCR4 tropic variants are the result of evolution from CCR5 viruses after transmission, or whether they are transmitted alongside CCR5 tropic variants (Archer, Rambaut et al. 2010; Bunnik, Swenson et al. 2011; Chalmet, Dauwe et al. 2012). This study found only one CXCR4 tropic subpopulation out of 238 subpopulations from the 36 individuals at an FPR of 5.0%. The subpopulation was the major subpopulation from the first time point (date of diagnosis) of a patient that was diagnosed very early after infection, prior to HIV antibody seroconversion. By the second time point, 8 days later, a single nucleotide mutation had resulted in the switch of an amino acid in the V3 loop, changing the prediction of tropism. For the remaining patients with clear growth of minor subpopulations over time points, there

was no evidence that the tropism of the virus as determined by the V3 loop region was instrumental in the process – though other aspects of the envelope gene itself both within and outside of the 400bp *env* amplicon may have influenced viral fitness (Lynch, Shen et al. 2009).

Another potential factor involved in evolution and competition between viruses that can be inferred directly from the deep-sequencing data is antiretroviral drug resistance. Though all individuals in the study were treatment naive, it is possible that some were infected by individuals failing therapy, and who had developed drug resistant viral strains as a consequence. Alternatively, a treatment-naive source may have carried transmitted drug resistance. It has been shown that the fitness costs imposed by drug resistant mutations in the absence of drug pressure leads to the loss of the drug resistant viral strain and reversion to wildtype virus archived prior to therapy initiation (Hedskog, Mild et al. 2010). However, there was insufficient evidence of drug resistant virus playing a role in the dynamics of the patients in this study, as only one of the 6 patients with a drug resistance mutation had multiple time points available. For this patient, patient 2, the revertant mutation T215S (García-Lerma, Nidtha et al. 2001) was present in all four major time points as the top subpopulations (>94% abundance) over a 25 week period.

There are a number of limitations to the study that need to be acknowledged. Although a useful window has been opened onto the complexities of subpopulation dynamics within single patients, and transmission of these subpopulations between patients, by the use of two different regions of the HIV genome, the limited nature of the genome coverage means that all regions of interest could not be investigated for possible roles in quasi-species evolution and interchange of primary subpopulations. Other studies that have carried out full genome deep-sequencing (some using 'deep-cloning') analysis have focused primarily on detecting epitopes important for targeting by the host immune pressure, with a view to

applying the findings to vaccine development (Keele, Giorgi et al. 2008; Henn, Boutwell et al. 2012), and have not attempted to detect large scale genomic rearrangements that might be driven by other factors affecting virus replication, such as cell entry mechanisms or virion release. This study focused on a subset of patients with recent infection, verified by laboratory and clinical information. The actual date of infection was estimated using a tool based upon Hamming distance frequency distributions (Giorgi, Funkhouser et al. 2010). Comparison of the results obtained for individuals with >1 *pol* subpopulation when mixed subpopulation reads (from subpopulations of different subtype) and primary subpopulation reads were submitted revealed that large differences can be observed in the estimated time to MRCA if the input is not corrected for infection by multiple founder virions that are genetically divergent (data not shown). Ideally, it would have been useful to have had more accurate information on time of infection for all patients, and to have caught them in the very early stages, such as for patient 8, sampled in Fiebig stage II. However, the average time to MRCA predicted by the Poisson-Fitter tool using only reads from the primary subpopulation was a median of 139 days (IQR: 125.5-156 days), and this matches well with the dates estimated by the guanidine based avidity assay initially employed to select for patients highly likely to have recently seroconverted (the avidity assay cut-off used in this study (≤0.75) identifies patients likely to have seroconverted within the previous 125 days (95 CI: 85- 164 days)). Finally, although the number of patients involved in this study is high in comparison to many previous HIV deep-sequencing studies (Hedskog, Mild et al. 2010; Bunnik, Swenson et al. 2011; Gianella, Delport et al. 2011), it is still smaller than ideal in terms of applicability of the findings to the general infected population. Selecting patients with sufficiently high viral loads to enable confident sampling of low level quasi-species prevalence imposes a major restriction on patient samples available for analysis, and means that the study, together with those previously highlighted, provides more of an illustration of the possible avenues of

evolution explored by HIV within and between patients, and how these can be radically different between individuals, suggesting possible areas for more in depth focus in future work.

In conclusion, this study has emphasised the dynamic nature of early phase HIV infection in a large cross-sectional sample of newly diagnosed patients, 9 with multiple time points allowing for analysis of quasi-species evolution. The study took a subpopulation reconstruction approach in order to treat mutations in different parts of the sequenced amplicons as linked, not as isolated variants that would have overlooked the interdependence of different regions of the HIV genome. In terms of evidence for infections being founded by single virions, the results of the clustering approach is in broad agreement with previous work (Keele, Giorgi et al. 2008; Fischer, Ganusov et al. 2010), with 30.6% of the 36 patients having evidence of >1 subpopulation in *pol*. The percentage is higher in *env*, at 47.2%, but this region of the viral genome is under greater host immune system pressure, and hence may evolve at a faster rate over the initial infection. The study also illustrates the utility and limitations involved in using single regions of the HIV genome for molecular epidemiological inference: a phylogenetically plausible transmission cluster detected using standard *pol* Sanger consensus sequence taken for routine drug resistance surveillance was also detected with molecular epidemiology of deep-sequencing data, but it was only in the deep-sequencing data from *pol* and *env*, and over multiple time points, that a more detailed, complex picture of the inter- and intra-patient evolutionary dynamics was revealed.

Chapter Six

Molecular epidemiology of new HIV-1 diagnoses in Kumasi, Ghana: a resource limited, generalised HIV-1 epidemic setting

6 Molecular epidemiology of new HIV-1 diagnoses in Kumasi, Ghana: a resource limited, generalised HIV-1 epidemic setting

Abstract

This chapter investigates a HIV-1 epidemic in the context of a resource-limited setting using a molecular epidemiological approach to investigate the extent to which the epidemic is structured into transmission clusters. The chapter also investigates the use of a nucleotide consensus sequence classifier of infection length developed in chapter three in assessing the extent to which individuals diagnosed with recent infection play a role within the cohort in question.

Newly HIV diagnosed individuals attending the Komfo Anokye HIV clinic in Kumasi, Ghana, had their blood sampled and shipped to the UK. The *pol* and *env* regions of the virus were sequenced, and an antibody avidity assay performed on patient samples to assess stage of infection. Phylogenetic analysis did not reveal extensive clustering of transmissions (4 transmission pairs out of 288 individuals with *pol* sequence available for analysis). Results from the avidity assay results and nucleotide ambiguity classifier did not correlate, but did co-segregate (high avidity/high nucleotide ambiguity, low avidity/low nucleotide ambiguity).

There does not appear to be a major clustering of transmissions within this setting, suggesting risk groups associated with clustering in resource-rich settings are not a major factor in Kumasi. Results from the application of the avidity assay and nucleotide ambiguity classifier suggest that both approaches may need to be further optimised for non-UK settings, but may prove useful in the future.

6.1 Introduction

In Ghana, a country with a population of 25.0 million (http://data.un.org), there were an estimated 240000 individuals living with HIV (type 1 and 2) in 2012, and the estimated HIV prevalence was 1.4% among those aged 15-49 years (UNAIDS 2013). Ghana is defined as a lower middle income country by the World Bank (http://data.worldbank.org/country/ghana), with a life expectancy at birth of 61 years (2011) and gross national income per capita of $1,550 (current US$, 2012) (c.f. United Kingdom: 81 years and $38,670 respectively). Antiretroviral therapy (ART) was first introduced in 2003 (Ohene and Forson 2009), and there are now 66366 (60.3%) HIV positive individuals receiving therapy, of an estimated 110000 individuals eligible for treatment (UNAIDS 2013). According to national guidelines, ART is initiated when patients have a $CD4^+$ cell count <350 cells/mm^3 and/or symptoms corresponding to WHO clinical stages 3 and 4 (National HIV/AIDS/STI Control Programme 2010). Monitoring of ART efficacy is carried out by assessing body weight, clinical events (e.g., opportunistic infections), $CD4^+$ cell counts, haemoglobin and serum biochemistry for hepatic transaminases and creatinine levels. HIV-1 RNA load monitoring is not routine, but is recommended to take place 6 months after commencement of therapy and subsequently every 12 months. Additionally, viral load monitoring may be requested in order to confirm treatment failure if the patient $CD4^+$ cell count is found to be declining, or clinical disease is progressing. The lack of viral load monitoring means that individuals failing ART are not identified early enough to limit emergence of drug resistance (Hamers, Wallis et al. 2011). In addition, limited second and third-line regimen options often result in clinicians being forced to allow patients to persist on failing therapy or recycle previously used agents (Bennett, Myatt et al. 2008). These considerations point to the development of a reservoir of resistant strains in treated people, posing a risk of onward transmission of resistant variants; indeed,

there is evidence of an increase in transmitted drug resistance since the roll-out of ART in west Africa (Gupta, Jordan et al. 2012; Stadeli and Richman 2013).

Hitherto, most molecular investigations of the HIV epidemic in Ghana have been based upon limited numbers of samples (Ishikawa, Janssens et al. 1996; Brandful, Ampofo et al. 1998; Sagoe, Dwidar et al. 2007), or have used samples obtained primarily prior to the roll out of ART (Fischetti, Opare-Sem et al. 2004). These studies have demonstrated a preponderance of CRF02_AG strains in the national epidemic. More recently, a larger molecular epidemiological study of 207 Ghanaian HIV-1-seropositive persons accessing the national ART program between 2002 and 2004 (Delgado, Ampofo et al. 2008) detected unique *pol* recombinant forms in 25% of the subjects, mostly derived from CRF02_AG. This suggests that multiple HIV-1 infections – occurring as either dual- or super-infections – were common. While two ART-naïve patients were found to harbour transmitted drug resistance (TDR) mutations in reverse transcriptase, the study was performed early after the start of the national ART programme in Ghana, and before TDR would have been expected to start to emerge (Frentz, Boucher et al. 2012; Stadeli and Richman 2013).

This study has focussed on one catchment area, namely patients newly diagnosed with HIV in 2008-2012 who attended a large teaching hospital in Kumasi, the second largest city of Ghana. The clinic catchment area is peri-urban, and over 5000 patients are in follow-up (every 3-4 months), with ~2500 having started ART. A 2007 study established that the majority of patients on ART in this cohort were female (79.7%), and that over a third of patients (34.8%) self-identified as unemployed (Ohene and Forson 2009). Our aim was to characterise the molecular epidemiology of HIV-1 infection in this regional cohort, and test the performance of the tools developed within a UK cohort for investigating time of infection and transmission patterns. Linked to this, one important aim of the study was to determine the prevalence of transmitted drug resistance within the now mature ART programme.

6.2 Methods

6.2.1 Study population

Serum and plasma samples were collected from randomly selected newly diagnosed patients attending the HIV clinic at Komfo Anokye Teaching Hospital in Kumasi, Ghana, between 2008 and 2012, stored locally at -80^{o}C, and shipped on dry ice to the UK for testing. Associated clinical and laboratory data were collected from the case records and anonymised. The study was approved by the Committee on Human Research Publications and Ethics at Kwame Nkrumah University of Science and Technology (KNUST), Kumasi, Ghana.

6.2.2 Guanidine-based avidity assay

Patient plasma or serum were defrosted and vortexed, prior to undergoing guanidine-based avidity testing on the Vitros ECiQ Immunodiagnostic System as previously described (section 2.3). An avidity index of ≤0.75 was regarded to be indicative of seroconversion within the previous 125 days (95 CI: 85 to 164 days) (Chawla, Murphy et al. 2007). Result interpretation also took into account the $CD4^{+}$ cell count at the time of presentation to reduce the risk of falsely assigning the infection to the recent category in patients with advanced immune compromise (Chargelegue, Stanley et al. 1995). Previous work associated with the UK cohort used to develop the nucleotide ambiguity cut-off (chapter 3) found that median $CD4^{+}$ cell counts for 103 individuals carefully defined as recent infections (within 125 days of seroconversion) was 565 cells/mm^3 (IQR: 415-760 cells/mm^3). For 101 patients from the same cohort with avidity indices ≤0.80 but with good evidence of an established infection, the median $CD4^{+}$ cell count was 121 cells/mm^3 (IQR: 25-357 cells/mm^3). Finally, the median $CD4^{+}$ cell count for 32 patients diagnosed with strong evidence of advanced disease/AIDS (e.g. PCP) was 42.5 cells/mm^3 (17-89 cells/mm^3). With these $CD4^{+}$ cell count medians and inter-quartile ranges in mind, cut-offs of 150 and 100 cells/mm^3 were selected as filters for

individuals with avidity indices ≤0.75, to gauge the effect of removing such individuals from the Kumasi new diagnosis cohort had on the overall proportion of patients classified as recently infected using both the avidity indices or nucleotide ambiguity cut-off.

6.2.3 Sequencing

Pol consensus sequencing was carried out as per a previously described in-house assay (section 2.5). A protocol was also developed to enable sequencing of a large region of the HIV-1 genome, including Rev, Vpu, gp120 and over half of gp41, in order to improve the molecular epidemiological analysis, as previously described (sections 2.6 and 2.7). The protocol followed a nested PCR strategy, with both sets of primers designed within genomic regions conserved across multiple subtypes of HIV-1 (A1, B, C, D, F, G, J, CRF01_AE, CRF02_AG and CRF06_cpx) in order to maximise the chances of successful amplification of novel subtypes and recombinant structures. Subtyping was carried out using the COntext-based Modelling for Expeditious Typing (COMET) web tool (http://comet.retrovirology.lu/).

6.2.4 Transmitted drug resistance

Pol consensus sequence electropherograms were manually inspected for quality prior to being submitted to the Calibrated Population Resistance tool (Gifford, Liu et al. 2009) for analysis using the World Health Organization 2009 Mutation List (Bennett, Camacho et al. 2009).

6.2.5 Transmission cluster identification

In order to identify any putative transmission clusters, *pol* sequence electropherograms were inspected for quality, aligned and stripped of drug resistance associated codons (Bennett, Camacho et al. 2009). Phylogenetic analysis was carried out using PhyML3.0 (GTR model of nucleotide evolution), with 1000 bootstrap trees generated using the FastTree2.1 program.

Potential transmission clusters were identified using criteria of ≥95% bootstrap support and maximum intra-cluster Hamming distance of ≤0.045.

6.3 Results

6.3.1 Study population and guanidine-based avidity assay results

A total of 456 subjects entered the study between 2008 and 2012 with median age 37yrs (IQR: 30-44yrs); 333 (73.0%) were female. Avidity testing was performed on samples taken within median 4 days (IQR: 1-68 days) of first HIV positive test. Overall 155/456 (34.0%) patients had an avidity index ≤0.75, and 301 (66.0%) had an index >0.75. $CD4^+$ cell counts of <150 and <100 cells/mm^3 at diagnosis were investigated as proxies for advanced disease stage, which increases the risk of falsely categorising an infection as recent. Among the 155 patients with avidity index ≤0.75, 13 had $CD4^+$ cell counts <150 and 4 <100 cells/mm^3 (with 8 individuals without baseline $CD4^+$ cell count results). Based on these criteria 143/452 (31.6%) and 134/443 (30.2%) patients showed evidence of a recent infection at diagnosis.

KATH HIV clinic new diagnosis cohort (n=260)	
Median age in years (IQR)	37 (30, 45)
Median CD4 cell count, cells/ml (IQR)	369 (225, 530)
Female (%)	181 (69.6)
Male (%)	79 (30.4)
Subtype (%)	
A1	6 (2.3)
B	1 (0.4)
C	1 (0.4)
D	1 (0.4)
G	7 (2.7)
CRF02_AG	204 (78.5)
CRF06_cpx	14 (5.4)
Other	26 (10.0)
Resistance (%)	
Any major mutation	7 (2.7)
Major NRTI mutations	3 (1.2)
Major NNRTI mutations	4 (1.5)
PIs (major mutations)	1 (0.4)
≥ 2 classes	1 (0.4)

Table 6.1. Demographic and clinical breakdown of the 260 patients for whom both clinical and sequence data were available.

6.3.2 HIV-1 subtypes

Overall, 288/456 (63.2%) plasma samples yielded a *pol* sequence. A total of 260 new diagnosis patients recruited from the HIV clinic had clinical and sequence data available. They were predominantly female (69.6%), with a median age of 37 (IQR: 30,45) and median $CD4^+$ cell count of 369 cells/ml (IQR: 225,530) (Table 6.1.). Based on *pol* sequences, CRF02_AG was the most common genetic type (225/288, 78.1%); the second most prevalent genetic type was CRF06_cpx (14, 4.9%). A total of 31 (10.8%) sequences could not be assigned to a recognised genetic type by the COMET algorithm, possibly representing unique

recombinant forms (URFs) (Table 6.2.). Of these, 12/31 *pol* sequences contained a combination of CRF02_AG and CRF06_cpx: the *env* regions were pure CRF02_AG in 2/12, pure CRF06_cpx in 4/12, and a mosaic of CRF02_AG and CRF06_cpx in 3/12 (Table 6.3.). To briefly investigate the extent to which subtyping results were affected by methodology and subtype reference sequences, all 288 *pol* sequences were run through the REGA subtyping algorithm for comparison with COMET and 34/288 (11.8%) sequences were found to have discordant results (23 were assigned CRF02_AG in COMET, but could not be assigned by REGA).

Pol subtype	Frequency
unassigned_1;02_AG, A1	2
unassigned_1;02_AG, D	1
unassigned_1;09_cpx, A1	1
unassigned_1;43_02G, G	1
unassigned_2;02_AG, 06_cpx	4
unassigned_2;02_AG, 09_cpx	2
unassigned_2;02_AG, A1	7
unassigned_2;02_AG, G	1
unassigned_2;06_cpx, 02_AG	8
unassigned_2;09_cpx, 45_cpx, 02_AG	1
unassigned_2;25_cpx, 02_AG, A1	1
unassigned_2;A1, 01_AE, 02_AG	1
unassigned_2;D, 02_AG	1

Table 6.2. A total of 31 *pol* sequences were unable to be unambiguously assigned to a subtype or CRF by the COMET subtyping tool. The COMET suggested structure is indicated after the semi-colon. The right hand column lists the number of sequences belonging to each structure.

	Potential CRF02_AG/CRF06_cpx recombinants			
	Pol subtype	Pol sequence length (bp)	Env subtype	Env sequence length (bp)
Seq 1	unassigned;06_cpx, 02_AG	1248	02_AG	2380
Seq 2	unassigned;06_cpx, 02_AG	1248	02_AG	2387
Seq 3	unassigned;06_cpx, 02_AG	1248	06_cpx	1386
Seq 4	unassigned;06_cpx, 02_AG	1246	06_cpx	2227
Seq 5	unassigned;02_AG, 06_cpx	1248	06_cpx	2210
Seq 6	unassigned;06_cpx, 02_AG	1248	06_cpx	2280
Seq 7	unassigned;02_AG, 06_cpx	1248	unassigned;02_AG, 06_cpx	2368
Seq 8	unassigned;06_cpx, 02_AG	1248	unassigned;02_AG, 06_cpx, 09_cpx	1575
Seq 9	unassigned;06_cpx, 02_AG	1248	unassigned;06_cpx, 02_AG	2027
Seq 10	unassigned;02_AG, 06_cpx	1248	unassigned;06_cpx, 02_AG	2297
Seq 11	unassigned;02_AG, 06_cpx	1248	unassigned;06_cpx, 02_AG, G	1381
Seq 12	unassigned;06_cpx, 02_AG	1248	unassigned;06_cpx, 02_AG, G	2370

Table 6.3. Table showing the *pol* and *env* region subtype/CRF assignment by COMET for the 12 viruses that appeared to be potential CRF02_AG/CRF06_cpx recombinants; 02_AG = CRF02_AG, 06_cpx = CRF06_cpx, 09_cpx = CRF09_cpx.

6.3.3 Transmitted drug resistance

288 sequences were submitted to the Calibrated Population Resistance tool (Gifford, Liu et al. 2009), to detect the presence of any drug resistance mutations listed on the World Health Organization 2009 Mutation List. Mutations detected were verified by checking the original electropherogram traces. Among 271 RT sequences suitable for analysis, 4 (1.5%) and 5 (1.8%) showed NRTI and NNRTI drug resistance associated mutations, respectively. There were two subjects (0.7%) with resistance to both drug classes. 271 sequences were suitable for analysis of protease, with 1 (0.4%) sequence having a drug resistance associated mutation. In total, 8/271 (3.0%) of patients in the cohort had a drug resistance mutation in their HIV-1 RT and protease sequence. Overall the results were consistent with drug availability in Ghana, comprising primarily NRTIs and NNRTIs (Table 6.4).

	Year diagnosed	NRTI	NNRTI	PI
Patient A	2008	D67N		
Patient B	2008	D67N		
Patient C	2009		V106A, G190A	
Patient D	2010		K103N	
Patient E	2011			I85V
Patient F	2011		K101E	
Patient G	2011	M184V	K103N	
Patient H	2011	M184V	K103N, Y181C	

Table 6.4. Drug resistant mutations by year of patient diagnosis.

6.3.4 Transmission clusters

288 patients had their HIV-1 *pol* consensus sequences subjected to phylogenetic analysis using PhyML3.0. Four pairs of sequences were identified as potential transmission clusters on the basis of having ≥95% bootstrap support and maximum intra-cluster pairwise genetic of ≤0.045 (Fig 6.1. and Table 6.5.). Consensus sequencing of the envelope region was carried out for all 8 patients to see if there was additional support for relatedness of the sequences. Percentage identity between *env* sequences within each pair was 93.1% or more (Table 6.4.), compared to a mean pairwise genetic identity of 88.0% when the pairwise distances between non-related pairs were used as a control. In all pairs there was good agreement between *pol* and *env* subtypes between the individuals in the pair and between regions within each individual.

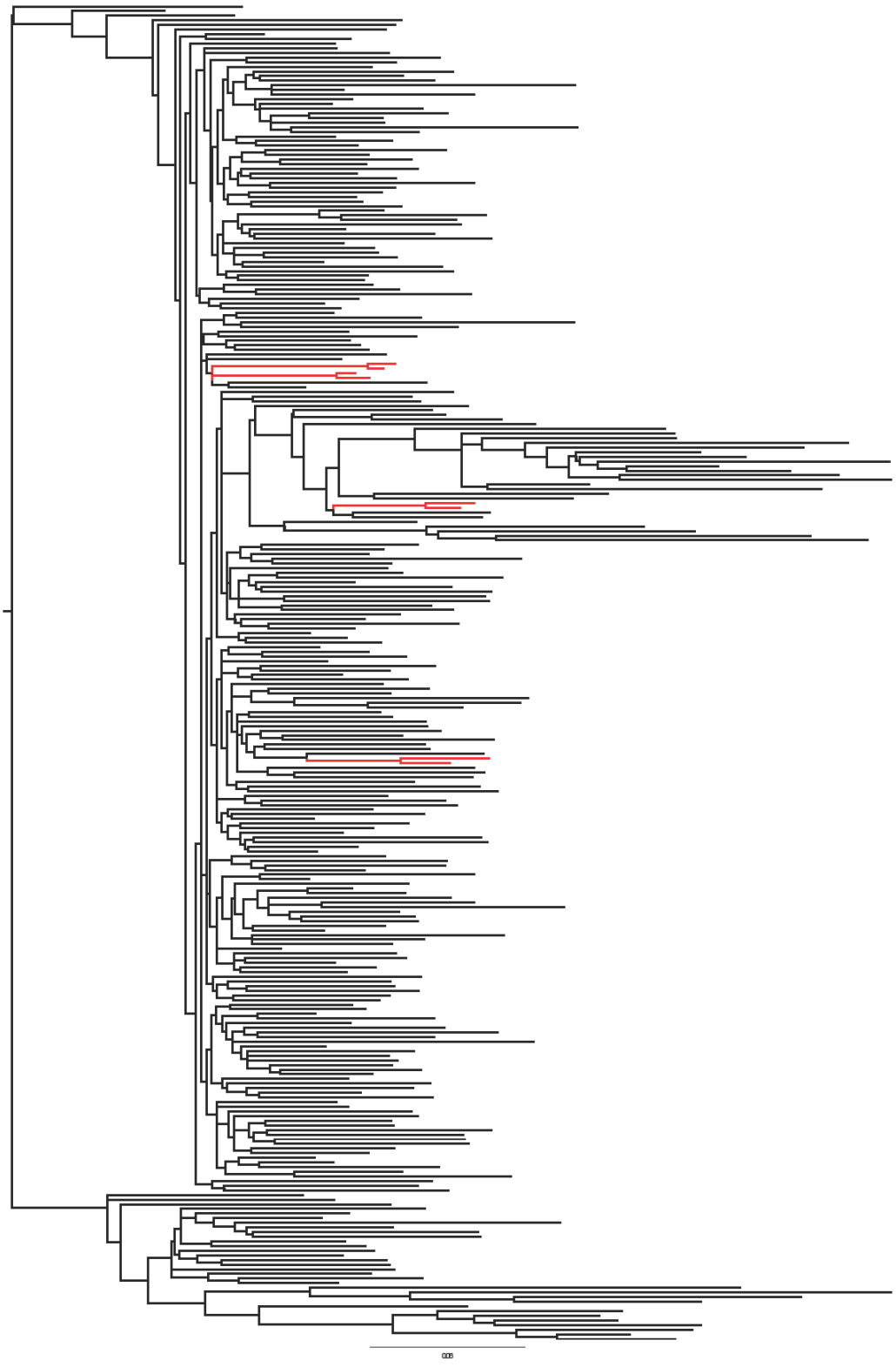


Figure 6.1. Phylogenetic tree showing four putative transmission clusters in red, as defined by ≥95% bootstrap and ≤0.045 intra-cluster pairwise genetic distance. Scale bar is 0.06 nucleotide substitutions per site.

	Date of diagnosis	Age at diagnosis	Gender	Baseline CD4	Pol % identity	Pol subtype	Env % identity	Env subtype
Pair 1	09 July 2008	22	F	342	97.9	CRF02_AG	96.1	CRF02_AG
	08 June 2008	25	M	419		CRF02_AG		CRF02_AG
Pair 2	18 August 2008	25	F	513	98.2	CRF02_AG, A1	93.9	CRF02_AG
	07 October 2011	32	F	362		CRF02_AG, A1		CRF02_AG
Pair 3	22 November 2010	47	F	401	96.1	CRF02_AG	94.2	CRF02_AG
	22 November 2010	54	M	171		CRF02_AG		CRF02_AG
Pair 4	28 November 2008	37	F	533	96.9	CRF02_AG	93.1	CRF02_AG
	12 April 2008	50	M	236		CRF02_AG		CRF02_AG

Table 6.5. Table showing clinical characteristics of the eight individuals involved in four putative transmission clusters.

6.3.4.1 Nucleotide ambiguity cut-off

The proportion of ambiguous nucleotides was calculated as previously described (section 2.12.2) for all 288 patients with *pol* sequence available to see if it could be a useful marker of infection length when correlated with avidity indices. After removing DRAM associated codons (Bennett, Camacho et al. 2009), 17/274 (6.2%) patients had 0.00% nucleotide ambiguities in their DRAM codon stripped *pol* sequence (14 patients removed due to sequence length <800bp, minimum length required for 0.00% cut-off), 37/288 (12.8%) patients had ≤0.17% nucleotide ambiguities in their *pol* sequence (as per the cut-off developed on a UK cohort of new diagnosis patients – see chapter 3). As advanced HIV disease is associated with a reduction in intra-patient quasi-species diversity (Shankarappa, Margolick et al. 1999), the nucleotide ambiguity analysis was repeated removing patients potentially diagnosed with late stage disease, using $CD4^+$ cell count as a proxy marker. In a similar procedure to the avidity indices analysis above, patients were ordered by avidity indices, and individuals with avidity indices ≤0.75 were filtered on the basis of $CD4^+$ cell counts <150 and <100 cells/mm^3. 17/267 (6.4%) and 17/272 (6.3%) had 0.00% ambiguous nucleotides using the <150 and <100 cells/mm^3 filters respectively (with sequences <800bp in length removed). 37/281 (13.2%) and 37/286 (12.9%) individuals had a percentage of ambiguous nucleotides <0.17% using the <150 and <100 cells/mm^3 filters respectively. Extending the $CD4^+$ cell count filter upwards to <250 and <350 cells/mm^3 did not change the results for either cut-off in a significant way (Table 6.6).

	Nucleotide ambiguity cut-off	
CD4 cut-off	**0.00%**	**0.17%**
100	17/272 (6.3)	37/286 (12.9)
150	17/267 (6.4)	37/281 (13.2)
200	17/263 (6.5)	37/277 (13.4)
250	17/260 (6.5)	33/260 (12.7)
300	17/254 (6.7)	33/254 (13.0)
350	17/247 (6.9)	30/247 (12.1)

Table 6.6. Table showing the effect of $CD4^+$ cell count filter on proportion of Kumasi cohort defined as recently infected using the nucleotide ambiguity cut-off (both the 0.00% cut-off and the 0.17% cut-off).

The ambiguous nucleotide percentages of the individuals with avidity indices ≤0.75 was compared to those of the population of individuals with avidity indices >0.75 using the Mann-Whitney test. The median percentage of ambiguous nucleotides for individuals with avidity indices ≤0.75 vs. >0.75 avidity was 0.54 vs. 0.80 (z = 3.71258, p-value <0.001) (14/288 patients with *pol* sequence did not have avidity results available). When patients with <150 $CD4^+$ cells/mm^3 were removed from the group of individuals with avidity indices ≤0.75, the median percentage of ambiguous nucleotides for individuals with avidity indices ≤0.75 vs. >0.75 was 0.46 vs. 0.80 (z = 3.98748, p-value <0.001, two-tailed test). When removing patients with <100 $CD4^+$ cells/mm^3 the respective values were 0.46 vs. 0.80 (z = 3.9925, p-value <0.001, two-tailed test). When nucleotide ambiguity was used to order patients, a nucleotide ambiguity cut-off of 0.63% was found to maximise the difference between the median avidity index of patients below this cut-off when compared to patients above this cut-off (0.77 vs. 0.91, z = 5.46769, p-value < 0.001, two-tailed test).

Correlation of avidity indices with nucleotide ambiguity score was poor regardless of removal of individuals with low $CD4^+$ cell count (<350 cells/mm^3: R^2 = 0.055; <250

cells/mm^3: R^2 = 0.052; <150 cells/mm^3: R^2 = 0.037; All: R^2 = 0.035). When avidity index ≤0.75 was used as the definition of recent infection, application of the nucleotide ambiguity cut-off produced positive predictive values (PPVs) across the four CD4$^+$ cell count filtering categories of <350 cells/mm^3: 0.38; <250 cells/mm^3: 0.46, <150 cells/mm^3: 0.48 and All: 0.48 (Table 6.7.).

All

	≤ 0.75	> 0.75	
≤ 0.17	12	13	25
> 0.17	37	108	145
	49	121	170

Sensitivity	0.24
Specificity	0.89
PPV	0.48
NPV	0.74

< 150 cells/mm3 removed

	≤ 0.75	> 0.75	
≤ 0.17	12	13	25
> 0.17	32	108	140
	44	121	165

Sensitivity	0.27
Specificity	0.89
PPV	0.48
NPV	0.77

< 250 cells/mm3 removed

	≤ 0.75	> 0.75	
≤ 0.17	11	13	24
> 0.17	26	108	134
	37	121	158

Sensitivity	0.30
Specificity	0.89
PPV	0.46
NPV	0.81

< 350 cells/mm3 removed

	≤ 0.75	> 0.75	
≤ 0.17	8	13	21
> 0.17	19	108	127
	27	121	148

Sensitivity	0.30
Specificity	0.89
PPV	0.38
NPV	0.85

Table 6.7. Sensitivity and specificity of the nucleotide ambiguity cut-off, using the guanidine-based avidity assay as the reference marker for recent infection. Each of the four panels reflects use of a different CD4$^+$ cell count cut-off to remove individuals from the recent infection group (as classified by an avidity index ≤0.75) in order to assess the possible impact of individuals diagnosed with advanced HIV. Positive predictive value

(PPV) and negative predictive values (NPV) are also shown, and were calculated as described in (Altman and Bland 1994).

6.4 Discussion

This study investigated a cohort of patients newly diagnosed with HIV-1 in Kumasi, Ghana, a resource limited, generalised HIV epidemic setting. It set out to investigate four main aspects of the HIV-1 epidemic: i) a simple assessment of subtype distribution in this region of Ghana; ii) the role of transmission clusters in the formation of the epidemic within the environs of Kumasi, Ghana's second largest city; iii) assessment of the level of drug resistance circulating within the patient population; and iv) the performance of a guanidine-based avidity assay and a *pol* consensus sequence nucleotide ambiguity cut-off in identifying patients in early stages of infection.

The results of the first part of the study fits well with previous more limited studies of populations elsewhere in Ghana: CRF02_AG was found to be the most prevalent form of the virus in the population, at 78.1% of sequences, followed by CRF06_cpx, at 4.9% (Fischetti, Opare-Sem et al. 2004; Sagoe, Dwidar et al. 2009). Both of these forms are known to be prevalent in this region, together with subtypes A and G. In comparison to other studies from Accra and the eastern region of Ghana (Delgado, Ampofo et al. 2008), the proportion of potential URFs of complex structure detected in this study is slightly lower. This may indicate that in and around Kumasi there are fewer individuals dually infected with HIV-1 of different forms, perhaps through lower levels of partner exchange or limited import of non-standard genetic forms of the virus compared to what might be occurring in the capital city or along particular trade routes. However, without larger sample numbers, more widespread population sampling, and full-genome sequencing, it may be futile to speculate in too much depth, and it should be noted that methodology and reference sequence sets (and how these reference sequences were themselves subtyped) can also have an impact on subtype analysis where complex genome structures are involved. As an example, the COMET subtyping tool uses a Prediction by Partial Match (ppm) algorithm with a reference set of 201 HIV-1

genomes, including CRFs 1-49, whereas the REGA HIV subtyping tool v2.0 (De Oliveira, Deforche et al. 2005) uses bootscanning, and has 48 reference sequences covering CRFs 1-14; consequently when the results for the 288 *pol* sequences for COMET were compared to REGA there were discordant results. The region of the genome submitted for analysis can also have a major impact on subtyping, as regions of a CRF that are exclusively of one subtype may disguise the true complexity of the genome (Sagoe, Dwidar et al. 2009), for example, in CRF02_AG the C2V3 region of *env* is exclusively subtype A, and so reliance solely on this region alone for subtyping is likely to lead to misclassification (Brandful, Ampofo et al. 1998). From the 31 potential URFs identified in this study, the fact that 12 of them have evidence of recombination between CRF02_AG and CRF06_cpx is not unexpected given that these are the major forms circulating in the region. However, if any particular URFs are found to be circulating in significant numbers, it may be interesting to perform whole-genome sequencing to map breakpoints and identify which segments of the genome have been retained from which parental strains, as this could point to important genetic factors involved in the evolution of HIV-1 in this population, i.e. factors that have an influence on transmission, or pathogenesis. It is important to note that in this cohort HIV is diagnosed using a rapid test that detects both HIV-1 and HIV-2 antibodies, but routine follow-up tests to differentiate between the two viruses are not performed. This means that samples that did not produce sequence may have been HIV-1 viruses of novel structure, with sequence that did not allow the reverse transcription primer to bind efficiently during the RT-PCR, or could have been HIV-2, which would not be amplified by the primers used in this study. The prevalence of HIV-2 in this cohort will need to be determined in order to more fully address this issue.

The second aspect of the study investigated the extent to which HIV-1 infections in the cohort cluster into groups of highly related viruses, which may indicate the existence of

transmission chains within the wider local epidemic. In any setting, caution must be taken when trying to identify transmission patterns within an epidemic, as sampling depth (i.e. the proportion of infected individuals captured within the study) and time taken for epidemiologically linked individuals to infect one another, and subsequently be diagnosed, can have major impacts upon results. In this study, only transmission pairs were identified, and although it is impossible to rule the involvement of additional individuals not captured by the clinic (indeed, such individuals must exist for one partner to have become infected in the first place), the epidemiological data for 3 of the pairs is strongly suggestive of transmissions within heterosexual partnerships. The patients involved in the remaining pair are both females and were diagnosed several years apart, with one possible scenario that could account for this being that a third individual, a male, could have either infected both females, or been infected by one female and then went on to transmit to the second female. For two of the four transmission pairs the date of diagnosis and date of sample were within 30 days, but both avidity indices and nucleotide ambiguity scores suggested both partners in each pair were unlikely to have had recent infection at diagnosis. The remaining two pairs had diagnosis dates and sample dates too far apart to allow for a valid application of either biomarker. Returning to the overall transmission cluster analysis, there do not appear to be large clusters of transmission in this cohort, which in other settings can be associated with high risk sexual or injecting drug use behaviour (Mathers, Degenhardt et al. 2008; Brenner, Roger et al. 2011).

The third aim of the study was to assess the prevalence of drug resistance within the ART-naive HIV-1 infected population. There is evidence of increasing prevalence of transmitted drug resistance in African countries where ART roll-out was commenced early last decade (Frentz, Boucher et al. 2012; Gupta, Jordan et al. 2012; Stadeli and Richman 2013). Antiretroviral therapy has been available in Ghana since 2003, but other recent studies

specifically in Ghanaian populations have not found high levels of drug resistance in treatment naive individuals (Delgado, Ampofo et al. 2008; Nii-Trebi, Ibe et al. 2013). The finding that 3.0% of individuals harbouring a drug resistance associated mutation in this cohort is in line with other countries in the region, and lower than is found in other settings (U.K. Collaborative Group on HIV Drug Resistance 2012; Bonney, Addo et al. 2013). Although ongoing population level monitoring is likely to be warranted given continuing use of ART and lack of routine virological monitoring, and the limited options available for first-line therapy in Ghana (National HIV/AIDS/STI Control Programme 2010), the level of drug resistance found in this treatment naive population, together with the lack of evidence of substantial clustering of infections, suggests transmitted drug resistance is not currently a major issue in this context. Instead, it may be more pertinent to focus on the treatment adherence of those who have already commenced therapy, as an earlier study of the Kumasi HIV infected population on treatment found that whilst most patients sampled (95.6%) were on triple therapy, as per national guidelines, self-reported adherence levels were 80.6%, and just over half of participants found the cost of the therapy unaffordable (Ohene and Forson 2009). Likely to be linked to these issues, evidence from a more recent study suggests that drug resistance in patients experiencing virological failure in this population is high (Nii-Trebi, Ibe et al. 2013).

The final aspect of this study aimed to investigate how two biomarkers of recent HIV-1 infection performed within a population who may have very different virus and host characteristics to the population the biomarkers were developed on. The biomarkers in question were a guanidine-based avidity assay, that measures the strength of HIV antigen-antibody affinity, and which was developed on a mixed subtype panel of infections in the United Kingdom (Chawla, Murphy et al. 2007), and a measure based upon the proportion of ambiguous nucleotides in *pol* consensus sequence, developed on a primarily Caucasian

population infected with HIV-1 subtype B (chapter 3). The proportion of patients diagnosed with recent infection based upon the avidity assay results (34.0%) was greater than the proportion predicted by both the 0.00% and 0.17% nucleotide ambiguity cut-off (6.2% and 12.8% respectively). Although the linear correlation between the two markers was poor, the statistically significant difference in median nucleotide ambiguity between patients above and below the 0.75 avidity index cut-off definition of recent infection is encouraging. Further, the fact that adjustment of the nucleotide ambiguity cut-off enabled two patient groups with a strong statistical difference in their avidity indices to be defined, indicates that the two markers do co-segregate well, and that both are able to detect some underlying substructure to the Kumasi cohort. The overall estimate of the proportion of recent infections in the cohort produced by the avidity assay compares with similar enzyme-immuno assays used in other African settings, however, these studies argue that such assays may overestimate the proportion of recent infections in a population (Karita, Price et al. 2007; Sakarovitch, Rouet et al. 2007; Braunstein, van de Wijgert et al. 2009; Kim, McDougal et al. 2010; Kim, Hallett et al. 2011). In comparison to the proportion suggested by the nucleotide ambiguity cut-off, this is the case (12.8% using the 0.17% cut-off), but further validation of both assays against a panel of known Ghanaian seroconverters is required to determine which assay is the more accurate and robust, and to determine sensitivity, specificity and positive predictive values. It may be that additional biomarkers of early infection, such as a combination of HIV antibody and p24 antigen tests (Fiscus, Pilcher et al. 2007), could be useful in combination with an enzyme-immuno assay, or that other sequencing-based infection length estimation methodologies, enabled by collection of dried blood or plasma spots, will soon prove tractable for cross-sectional surveillance purposes (Giorgi, Funkhouser et al. 2010; Ji, Li et al. 2011; Poon, McGovern et al. 2011; Dudley, Chin et al. 2012).

In conclusion, this study represents one of the largest analyses of the Human Immunodeficiency Virus in Ghana to date. Results of *pol* consensus sequence analysis confirms the broad distribution of HIV-1 genetic forms that have been found elsewhere in the country, over several years, and suggests an evolutionary status quo in terms of the predominant forms circulating, with a continuing turnover of novel recombination events occurring in a small proportion of individuals. Molecular epidemiological analysis of this cohort suggests that within this large peri-urban setting, there is not a major issue with high risk behaviour generating large chains of transmission, assuming reasonable sampling of the HIV infected population has been achieved by the clinic. The study also suggests that while transmitted drug resistance in treatment naive individuals does exist, it is at low levels, and resistance developed in individuals with virological breakthrough may therefore be more of an issue for clinicians, though this patient group was not a focus of this study. In terms of application of laboratory-based diagnostic assays designed to assess length of HIV-1 infection, results obtained in this study do compare with other studies investigating similar assays, but further work is likely required to ensure cut-offs are suitable for different populations in different settings.

Chapter Seven

General discussion

7 Discussion and future directions

The overall aims of this thesis were to investigate the molecular epidemiology and evolution of HIV-1 among newly infected patients in high and low prevalence settings. In the context of this thesis, those settings were the United Kingdom and Ghana. The investigation aimed to develop a measure of the length of HIV-1 infection based upon the number of ambiguous nucleotides in Sanger *pol* consensus sequences, and to then apply this measure to viral sequences obtained from cohorts of newly diagnosed patients. The work assessed the extent to which individuals with recent infections contribute to transmission clusters, and the levels of transmitted drug resistant viruses within the treatment naive population in both settings. In terms of the Ghana molecular epidemiological exploration, in light of the relative paucity of recent data available for large cohorts, this investigation also provided the chance to obtain a broader picture of the overall HIV-1 epidemic taking place in and around the country's second largest city, Kumasi. Finally, the investigation involved a more in-depth analysis of the intra-host dynamics of recent HIV-1 infection, using deep-sequencing to look beyond the limitations imposed by Sanger sequencing.

An initial step in the investigation was to develop a measure of HIV-1 infection length that could be used with viral sequences that are collected routinely as part of antiretroviral drug resistance surveillance programs in the UK and elsewhere. Collections of such sequences exist in clinics or central repositories in a variety of countries, and offer a useful resource for investigations into the evolution and continuation of the epidemic (Yerly, Kaiser et al. 1999; Lewis, Hughes et al. 2008; Brenner, Roger et al. 2011). Determining the main drivers of an epidemic has a clear impact on containment policies. In the case of HIV, three sources of onward transmission can be proposed:

i) Subjects with recent infection that may or not have been diagnosed at the time when they experience peak virus replication and infectivity. A prominent role would trigger interventions to increase access to HIV testing and promote awareness of risk in the community.

ii) Subjects with established but untreated infection. A prominent role would trigger recommendations about earlier initiation of antiretroviral therapy as a public health measure

iii) Subjects with established and treated infections that experience suboptimal HIV suppression. A prominent role would trigger monitoring and management interventions to promote improved use of and adherence to treatment.

There is support from a number of studies that individuals with recent infections disproportionately drive the HIV epidemic in a variety of different settings compared to untreated individuals in the established, symptomless phase of their infection (Yerly, Vora et al. 2001; Brenner, Roger et al. 2007; Abu-Raddad and Longini Jr 2008; Hollingsworth, Anderson et al. 2008). These studies are necessarily carried out on cohorts of restricted size, as it is usually laboratory and/or clinical approaches that are used to assess the stage of infection of an individual, and as such there is a limitation on how many individuals such information can be collected for. A classifier that can be applied to nucleotide sequences taken routinely from diagnosed individuals provides a potential way of getting around this issue. In this study, two classifier cut-offs were investigated, one that was optimised to maximise both sensitivity and specificity, and one that aimed to maximise specificity for recent infection at the expense of sensitivity. The validation and overall performance of both classifiers seemed strongly influenced by sample size, with results being relatively poor for the smaller St Mary's Hospital cohort, compared to the larger UK Register of HIV Seroconverters cohort. This is possibly an effect of the margin of error associated with

smaller sample sizes (Lohr 2009), and the fact that the prevalence of recent infection in these cohorts was relatively low, and so the number of false positive classifications may be inflated (Altman and Bland 1994). Comparison of the positive predictive values obtained for the 0.17% and 0.00% ambiguity cut-offs suggests that the more stringent 0.00% cut-off may be more useful in terms of identification of recent infections. The performance of both classifiers on the UKHIVDRD treatment naive cohort produced estimates of the proportion of recent infections in the population consistent with different studies in European settings (Semaille, Barin et al. 2007; Yerly, Junier et al. 2009; Fisher, Pao et al. 2010). The closest match in terms of geographical setting and definition of recent infection actually comes from results of the Recent Infection Testing Algorithm obtained from across England, Wales and Northern Ireland (Aghaizu, Brown et al. 2013). The results are from a chronologically more recent cross-section of the population, but suggest that overall 16% of new diagnoses were from the 4-6 month window post-infection. This result matches closely to the UKHIVDRD result of 14% when the 0.00% cut-off was used, and therefore lends support to the use of this cut-off, and adds validity to the findings of the subsequent phylogenetic analyses that utilise the cut-off results.

The results of these subsequent phylogenetic analyses suggest that recent infections are more likely to be linked to transmission clusters than established infections, a finding that is supported by, but extends the scale of similar studies (Yerly, Vora et al. 2001; Pao, Fisher et al. 2005; Brenner, Roger et al. 2007; Brenner, Roger et al. 2008; Fisher, Pao et al. 2010). What the results of this investigation reveal are that this disproportionate linkage is seen across multiple subtypes, which are likely to represent different, compartmentalised sub-epidemics within the UK (Fox, Castro et al. 2010; Abecasis, Wensing et al. 2013). The linkage was also observed across different cluster sizes, from pairs to clusters containing 10 individuals or more. In terms of this disproportionate linkage of recent infections to clusters,

treatment of the results requires caution on two fronts. Firstly, there is evidence to suggest that clustering of recent infections in phylogenies may be a reflection of the method of identification of these clusters, whereby genetically closely related sequences are more likely to be linked together than genetically distant sequences (Volz, Koopman et al. 2012). Further to this, the process of contact tracing is found to be more efficient in more clustered networks (Eames and Keeling 2003; Kiss, Green et al. 2005), which may combine with increased probability of recall of previous sexual partners in individuals recently infected and thereby lead to preferential attachment of subsequent recent infections to clusters. Conversely, this effect may be countered by increased inability to recall previous partners if that number is large (Brewer, Garrett et al. 1999; Brewer and Garrett 2001), and so it is clear that there is scope for further investigation of this potential source of ascertainment bias. Secondly, as outlined in Brown et al. (Brown, Gifford et al. 2009), linkage of recent infections to clusters is not necessarily evidence for transmission from the recent phase of infection, as the period is by definition transient – individuals may have their virus sampled during the recent phase of their infection, but may not transmit for another half a year, and depending on the criteria used to identify clusters, they may still be identified as belonging to a transmission cluster (providing linked individuals have also been sampled). The thesis attempted to develop two proof-of-principle methods to illustrate how results from the nucleotide ambiguity classifier could be harnessed in overcoming this issue. Both methods aimed to allow a comparison of the contribution of recent infections to onward transmission between different cluster sizes, rather than an absolute figure for the proportion of infections coming from the recent infection phase. The results of the models, based on overlapping approaches, suggest that larger clusters seem to have a greater degree of overlap between the recent phase windows of infection, i.e. there appears to be a greater degree of transmission from the recent phase of infection. Whilst other studies have already concluded this, this is the first phylogeny-based

model that explicitly attempts to tackle the potentially misleading conflation of recent infection linkage to transmission clusters, and recent phase transmission. Future work is needed to disentangle these results from the potential ascertainment biases outlined above, to reveal if recent infections are really leading to a disproportionate number of onward infections, or they are just more likely to be diagnosed and to be linked in clusters through contact tracing and phylogenetic effects, but the models are a good example of how the nucleotide ambiguity classifier can be used in such large-scale analyses. Beyond this, it may be useful to investigate the usefulness of incorporating the classifier into recent infection testing algorithms as an additional biomarker of recent infection.

The molecular epidemiological characterisation of the Kumasi cohort represents one of the largest studies of its kind in Ghana to date. Previous studies have found a broadly similar range of subtypes to the results in this thesis, with CRF02_AG and CRF06_cpx predominating, but these studies found a slightly greater proportion of unique recombinant forms (Fischetti, Opare-Sem et al. 2004; Delgado, Ampofo et al. 2008; Sagoe, Dwidar et al. 2009). As discussed in chapter 4, these differences can be influenced by methodology and reference sequences used for subtyping. What is clear from the results in this thesis, and the published data, is that CRF02_AG is the predominant form of the virus in multiple regions of Ghana, in line with the rest of west Africa, and there is evidence to support the hypothesis that CRF02_AG has higher replicative fitness compared to its parental strains, which may have assisted its spread through the region (Fischetti, Opare-Sem et al. 2004; Konings, Burda et al. 2006; Njai, Gali et al. 2006). Whilst this predominance of CRF02_AG seems to be stable, there is clearly ongoing recombination between CRF02_AG and other forms of the virus, such as CRF06_cpx, to form new unique recombinant forms (URFs) of the virus. We did not detect any putative new CRFs in the Kumasi cohort, although there is clear potential for spread of the multiple URFs detected by us and others. These results point to an overall

relatively stable picture of subtype and CRF distribution in the region, broadly in line with global trends leading up to the period during which these viral samples were obtained (Hemelaar, Gouws et al. 2011).

The investigation of the Kumasi cohort also aimed to characterise the number of individuals that are diagnosed with recent infection in the region, using two approaches developed and used in the UK setting, the avidity assay and the nucleotide ambiguity classifier. There did not seem to be a strong correlation between the avidity index and proportion of nucleotide ambiguity in the cohort, but it is difficult to identify the causes of the discrepancies without additional measures of timing of infection. However, it is interesting to note the statistically significant difference in nucleotide ambiguity between patients with an avidity index ≤ 0.75, and patients with an avidity index >0.75, suggesting there is some relationship between the markers, and that a different cut-off for classifying infections as recent by the avidity assay may be needed in this population, as put forward elsewhere (Sakarovitch, Rouet et al. 2007; Kim, McDougal et al. 2010). It is important to appreciate that the cut-off of 0.75 was developed within Western cohorts of acute seroconverters predominantly infected with subtype B. There is published evidence that the avidity maturation curve is similar in at least some non-B subtypes, but the evidence is not extensive (Chawla, Murphy et al. 2007). As the Kumasi new diagnosis cohort was recruited cross-sectionally, as opposed to prospectively, it was not possible to identify a panel of seroconverters of known infection date against which to validate the two methods in this geographical setting. These studies need to be performed if either of these approaches is intended to be rolled out as a surveillance tool in this or similar settings.

The results of the phylogenetic analysis and transmission cluster identification in the Kumasi cohort suggests that in this cohort there is not a large degree of high risk sexual behaviours or use of injecting drugs, which can be drivers in the formation of transmission

clusters in other contexts (Mathers, Degenhardt et al. 2008; Brenner, Roger et al. 2011). One potential caveat to this is that there may be a higher degree of smaller clusters that are not being picked up due to the depth of sampling of the underlying epidemic that has been achieved in this thesis, however, this is difficult to ascertain with any degree of accuracy whilst estimates for the proportion of undiagnosed infections are absent in Ghana.

The final part of this thesis used a deep-sequencing approach to dissect HIV-1 quasi-species dynamics close to the point of infection to reveal some important issues in relation to the use of the nucleotide ambiguity classifier in estimating length of HIV-1 infection. There were a number of individuals who appeared to have more than one subpopulation at their earliest time point, which manifested in Sanger sequencing of the *pol* gene as a high number of ambiguous nucleotides – these subjects would have been classified as having an established infection using the nucleotide ambiguity classifier. If the proportion of such individuals remained the same on the larger, national scale, this could lead to an underestimation of the number of individuals with recent infection using the nucleotide ambiguity classifier. On the other hand, it is also clear that genetic diversity can increase without an increase in subpopulation number, which presents an opposite problem when considering application of the nucleotide ambiguity classifier. These within subpopulation increases in diversity may not be large enough to register on a Sanger sequence electropherogram, and so will not be picked up as ambiguous nucleotides, potentially leading to individuals with established infection being falsely classed as having a recent infection. This problem is less likely to be an issue going forward, with application of deep-sequencing technologies (Giorgi, Funkhouser et al. 2010; Poon, McGovern et al. 2011). However, in order to utilise the vast amount of historical HIV-1 *pol* sequences available as part of national drug resistance surveillance programs, both the potential for false negative and false positive results based on nucleotide ambiguity need to be taken into account and further investigated,

in terms of potential ways to address these limitations, perhaps by looking for hyper-diversity at particular sites known to differ between subtypes, perhaps by incorporating other clinical markers where available.

A further interesting aspect of HIV-1 molecular epidemiology revealed by analysis of the deep-sequencing data produced for this thesis is the identification and dissection of a transmission cluster within the patient cohort. Strong phylogenetic evidence connected 4 individuals, and enabled examination of the possible complexity in transmission dynamics that can be missed with consensus Sanger sequencing, and sequencing of one region of the HIV-1 genome only. Three individuals were linked on the basis of Sanger sequencing of the virus *pol* region performed on their diagnosis sample. The deep-sequencing results made it clear that 21 weeks later, the primary and secondary subpopulations in one patient had switched, and that if *pol* Sanger sequencing had been carried out using this latter time point, one of the three patients would not have been identified as belonging to the transmission cluster. If HIV-1 superinfection is more common than previously thought (Redd, Mullis et al. 2012; Redd, Quinn et al. 2013), and switches in primary and secondary/tertiary subpopulations occur on these timescales, the true structure of epidemics has hitherto been over-simplified. In the future, whole genome deep-sequencing, with an approach tailored to detect such infections, will be required for a more accurate understanding of the public health intervention approaches required to reduce transmissions.

References

454 Life Sciences Corp. (2009). "Using Multiplex Identifier (MID) Adaptors for the GS FLX Titanium Chemistry - Basic MID Set." 454 Sequencing Technical Bulletin No. 004-2009.

Abecasis, A.B., A.M.J. Wensing, et al. (2013). "HIV-1 Subtype Distribution and its Demographic Determinants in Newly Diagnosed Patients in Europe Suggest Highly Compartmentalized Epidemics." Retrovirology 10(1): 7.

Abu-Raddad, L. J. and I. M. Longini Jr (2008). "No HIV stage is dominant in driving the HIV epidemic in sub-Saharan Africa." AIDS 22(9): 1055-1061.

Aggarwal, I., M. Smith, et al. (2006). "Evidence for Onward Transmission of HIV-1 Non-B Subtype Strains in the United Kingdom." Journal of Acquired Immune Deficiency Syndromes (1999) 41(2): 201.

Aghaizu, A., A. Brown, et al. (2013). "HIV in the United Kingdom 2013 Report: data to end 2012." Public Health England, London.

Altman, D. G. and J. M. Bland (1994). "Statistics Notes: Diagnostic tests 2: predictive values." BMJ 309(6947):102.

Ananworanich, J., J. L. K. Fletcher, et al. (2013). "A Novel Acute HIV Infection Staging System Based on 4th Generation Immunoassay." Retrovirology 10(1):56.

Anderson, D., J. A. Politch, et al. (2011). "HIV Infection and Immune Defense of the Penis." American Journal of Reproductive Immunology 65(3): 220.

Andersson, E., W. Shao, et al. (2013). "Evaluation of Sequence Ambiguities of the HIV-1 *pol* gene as a Method to Identify Recent HIV-1 Infection in Transmitted Drug Resistance Surveys." Infection, Genetics and Evolution 18:125.

Anonymous (1988). "Human Immunodeficiency Virus Infection in the United Kingdom: Quarterly Report I. The Epidemic to 30 September 1987." Journal of Infection 16:291.

Apetrei, C., I. Loussert-Ajaka, et al. (1996). "Lack of Screening Test Sensitivity During HIV-1 Non-Subtype B Seroconversions." AIDS 10(14):F57.

Archer, J., A. Rambaut, et al. (2010). "The Evolutionary Analysis of Emerging Low Frequency HIV-1 CXCR4 Using Variants Through Time—an Ultra-deep Approach." PLoS Computational Biology 6(12):e1001022.

Arhel, N. (2010). "Revisiting HIV-1 Uncoating." Retrovirology 7(1):96.

Ariën, K. K., A. Abraha, et al. (2005). "The Replicative Fitness of Primary Human Immunodeficiency Virus Type 1 (HIV-1) Group M, HIV-1 Group O, and HIV-2 Isolates." J Virol 79(14):8979.

Arts, E. J. and D. J. Hazuda (2012). "HIV-1 Antiretroviral Drug Therapy." Cold Spring Harbor Perspectives in Medicine 2(4).

Baccam, P., R. J. Thompson, et al. (2001). "PAQ: Partition Analysis of Quasispecies." Bioinformatics 17(1):16.

Balzarini, J., P. Herdewijn, et al. (1989). "Differential Patterns of Intracellular Metabolism of 2', 3'-Didehydro-2', 3'-Dideoxythymidine and 3'-Azido-2', 3'-Dideoxythymidine, Two Potent Anti-Human Immunodeficiency Virus Compounds." Journal of Biological Chemistry 264(11):6127.

Bar, K. J., H. Li, et al. (2010). "Wide Variation in the Multiplicity of HIV-1 Infection Among Injection Drug Users." J Virol 84(12):6241.

Barré-Sinoussi, F., J.-C. Chermann, et al. (1983). "Isolation of a T-lymphotropic Retrovirus from a Patient at Risk for Acquired Immune Deficiency Syndrome (AIDS)." Science 220(4599):868.

Bennett, D. E., R. J. Camacho, et al. (2009). "Drug Resistance Mutations for Surveillance of Transmitted HIV-1 Drug-Resistance: 2009 Update." PloS One 4(3):e4724.

Bennett, D. E., M. Myatt, et al. (2008). "Recommendations for Surveillance of Transmitted HIV Drug Resistance in Countries Scaling up Antiretroviral Treatment." Antiviral Therapy 13:25.

Berkhout, B., A. Gatignol, et al. (1990). "TAR-independent Activation of the HIV-1 LTR: Evidence That Tat Requires Specific Regions of the Promoter." Cell 62(4):757.

Bezemer, D., A. van Sighem, et al. (2010). "Transmission Networks of HIV-1 Among Men Having Sex with Men in the Netherlands." AIDS 24(2):271.

Boeras, D. I., P. T. Hraber, et al. (2011). "Role of Donor Genital Tract HIV-1 Diversity in the Transmission Bottleneck." Proceedings of the National Academy of Sciences 108(46):E1156.

Boggiano, C. and D. R. Littman (2007). "HIV's Vagina Travelogue." Immunity 26(2):145.

Boily, M.-C., R. F. Baggaley, et al. (2009). "Heterosexual Risk of HIV-1 Infection per Sexual Act: Systematic Review and Meta-analysis of Observational Studies." The Lancet Infectious Diseases 9(2):118.

Bonney, E. Y., N. A. Addo, et al. (2013). "Low Level of Transmitted HIV Drug Resistance at Two HIV Care Centres in Ghana: A Threshold Survey." Ghana Medical Journal 47(2):82.

Booth, C. L., A. M. Garcia-Diaz, et al. (2007). "Prevalence and Predictors of Antiretroviral Drug Resistance in Newly Diagnosed HIV-1 Infection." Journal of Antimicrobial Chemotherapy 59(3):517.

Booth, C. L. and A. M. Geretti (2007). "Prevalence and Determinants of Transmitted Antiretroviral Drug Resistance in HIV-1 Infection." Journal of Antimicrobial Chemotherapy 59(6):1047.

Bordería, A. V., R. Lorenzo-Redondo, et al. (2010). "Initial Fitness Recovery of HIV-1 is Associated with Quasispecies Heterogeneity and Can Occur Without Modifications in the Consensus Sequence." PloS One 5(4): e10319.

Boyd, A. E., S. Murad, et al. (2005). "Ethnic Differences in Stage of Presentation of Adults Newly Diagnosed with HIV-1 Infection in South London." HIV Medicine 6(2):59.

Brandful, J. A. M., W. K. Ampofo, et al. (1998). "Sequence Note: Genetic and Phylogenetic Analysis of HIV Type 1 Strains from Southern Ghana." AIDS Research and Human Retroviruses 14(9):815.

Braunstein, S. L., J. H. van de Wijgert, et al. (2009). "HIV Incidence in Sub-Saharan Africa: A Review of Available Data with Implications for Surveillance and Prevention Planning." AIDS Rev 11(3):140.

Brenchley, J. M., T. W. Schacker, et al. (2004). "CD4+ T Cell Depletion During All Stages of HIV Disease Occurs Predominantly in the Gastrointestinal Tract." The Journal of Experimental Medicine 200(6):749.

Brenner, B. G., M. Roger, et al. (2008). "Transmission Networks of DrugResistance Acquired in Primary/Early Stage HIV Infection." AIDS 22(18):2509.

Brenner, B. G., M. Roger, et al. (2007). "High Rates of Forward Transmission Events After Acute/Early HIV-1 Infection." Journal of Infectious Diseases 195(7):951.

Brenner, B. G., M. Roger, et al. (2011). "Transmission Clustering Drives the Onward Spread of the HIV Epidemic Among Men who have Sex with Men in Quebec." Journal of Infectious Diseases 204(7):1115.

Brewer, D. D. and S. B. Garrett (2001). "Evaluation of Interviewing Techniques to Enhance Recall of Sexual and Drug Injection Partners." Sexually Transmitted Diseases 28(11):666.

Brewer, D. D., S. B. Garrett, et al. (1999). "Forgetting as a Cause of Incomplete Reporting of Sexual and Drug Injection Partners." Sexually Transmitted Diseases 26(3):166.

Briggs, J. A. G., K. Grünewald, et al. (2006). "The Mechanism of HIV-1 Core Assembly: Insights from Three-dimensional Reconstructions of Authentic Virions." Structure 14(1):15.

Briggs, J. A. G., T. Wilk, et al. (2003). "Structural Organization of Authentic, Mature HIV-1 Virions and Cores." The EMBO Journal 22(7):1707.

Briz, V., E. Poveda, et al. (2006). "HIV Entry Inhibitors: Mechanisms of Action and Resistance Pathways." Journal of Antimicrobial Chemotherapy 57(4):619.

Brooks, J. T., K. E. Robbins, et al. (2006). "Molecular Analysis of HIV Strains from a Cluster of Worker Infections in the Adult Film Industry, Los Angeles 2004." AIDS 20(6):923.

Brown, A. E., R. J. Gifford, et al. (2009). "Phylogenetic Reconstruction of Transmission Events from Individuals with Acute HIV Infection: Toward More Rigorous Epidemiological Definitions." Journal of Infectious Diseases 199(3):427.

Brown, A. J., D. Lobidel, et al. (1997). "The Molecular Epidemiology of Human Immunodeficiency Virus type 1 in Six Cities in Britain and Ireland." Virology 235(1):166.

Brown, M. P. S., Y. Guo, et al. (2013). "Rapid Sequencing of HIV-1 Genomes as Single Molecules from Simple and Complex Samples." 20th Conference on Retroviruses and Opportunistic Infections.

Bukrinsky, M. I., N. Sharova, et al. (1993). "Association of Integrase, Matrix, and Reverse Transcriptase Antigens of Human Immunodeficiency Virus type 1 with Viral Nucleic Acids Following Acute Infection." Proceedings of the National Academy of Sciences 90(13):6125.

Bunnik, E. M., L. C. Swenson, et al. (2011). "Detection of Inferred CCR5-and CXCR4-using HIV-1 Variants and Evolutionary Intermediates Using Ultra-deep Pyrosequencing." PLoS Pathogens 7(6):e1002106.

Buonaguro, L., M. L. Tornesello, et al. (2007). "Human Immunodeficiency Virus type 1 Subtype Distribution in the Worldwide Epidemic: Pathogenetic and Therapeutic Implications." J Virol 81(19):10209.

Campsmith, M. L., P. H. Rhodes, et al. (2010). "Undiagnosed HIV Prevalence Among Adults and Adolescents in the United States at the End of 2006." Journal of Acquired Immune Deficiency Syndromes 53(5):619.

Carlsen, T., A. B. Aas, et al. (2012). "Don't Make a Mista(g)ke: is Tag Switching an Overlooked Source of Error in Amplicon Pyrosequencing Studies?" Fungal Ecology http://dx.doi.org/10.1016/j.funeco.2012.06.003.

Chalmet, K., K. Dauwe, et al. (2012). "Presence of CXCR4-using HIV-1 in Patients with Recently Diagnosed Infection: Correlates and Evidence for Transmission." Journal of Infectious Diseases 205(2):174.

Chalmet, K., D. Staelens, et al. (2010). "Epidemiological Study of Phylogenetic Transmission Clusters in a Local HIV-1 Epidemic Reveals Distinct Differences Between Subtype B and non-B Infections." BMC Infectious Diseases 10(1):262.

Chang, J. J. and M. Altfeld (2010). "Innate Immune Activation in Primary HIV-1 Iinfection." Journal of Infectious Diseases 202(Supplement 2):S297.

Chargelegue, D., C. M. Stanley, et al. (1995). "The Affinity of IgG Antibodies to Gag p24 and p17 in HIV-1-infected Patients Correlates with Disease Progression." Clinical & Experimental Immunology 99(2):175.

Chawla, A., G. Murphy, et al. (2007). "Human Immunodeficiency Virus (HIV) Antibody Avidity Testing to Identify Recent Infection in Newly Diagnosed HIV type 1 (HIV-1)-seropositive Persons Infected with Diverse HIV-1 Subtypes." Journal of Clinical Microbiology 45(2):415.

Christ, F. and Z. Debyser (2013). "The LEDGF/p75 Integrase Interaction, a Novel Target for anti-HIV Therapy." Virology 435(1):102.

Chun, T.-W., D. Engel, et al. (1998). "Early Establishment of a Pool of Latently Infected, Resting CD4+ T Cells During Primary HIV-1 Infection." Proceedings of the National Academy of Sciences 95(15):8869.

Chun, T.-W. and A. S. Fauci (2012). "HIV Reservoirs: Pathogenesis and Obstacles to Viral Eradication and Cure." AIDS 26(10):1261.

Clapham, P. R. and Á. McKnight (2001). "HIV-1 Receptors and Cell Tropism." British Medical Bulletin 58(1):43.

Clavel, F. and A. J. Hance (2004). "HIV Drug Resistance." New England Journal of Medicine 350(10):1023.

Clewley, J. P., C. Arnold, et al. (1996). "Diverse HIV-1 Genetic Subtypes in UK." The Lancet 347(9013):1487.

Clumeck, N., F. Mascart-Lemone, et al. (1983). "Acquired Immune Deficiency Syndrome in Black Africans." The Lancet 321(8325):642.

Coffin, J. M. (1995). "HIV Population Dynamics In Vivo: Implications for Genetic Variation, Pathogenesis, and Therapy." Science 267(5197):483.

Cohen, M. S., C. L. Gay, et al. (2010). "The Detection of Acute HIV Infection." Journal of Infectious Diseases 202(Supplement 2):S270.

Cohen, M. S., G. M. Shaw, et al. (2011). "Acute HIV-1 Infection." New England Journal of Medicine 364(20):1943.

Collier, A. C., R. W. Coombs, et al. (1996). "Treatment of Human Immunodeficiency Virus Infection with Saquinavir, Zidovudine, and Zalcitabine." New England Journal of Medicine 334(16):1011.

Connor, R. I., K. E. Sheridan, et al. (1997). "Change in Coreceptor Use Correlates with Disease Progression in HIV-1–infected Individuals." The Journal of Experimental Medicine 185(4):621.

Coombs, R. W., C. E. Speck, et al. (1998). "Association Between Culturable Human Immunodeficiency Virus Type 1 (HIV-1) in Semen and HIV-1 RNA Levels in Semen and Blood: Evidence for Compartmentalization of HIV-1 between Semen and Blood." Journal of Infectious Diseases 177(2):320.

Cornelissen, M., S. Jurriaans, et al. (2007). "Routine HIV-1 Genotyping as a Tool to Identify Dual Infections." AIDS 21(7):807.

Craigie, R. and F. D. Bushman (2012). "HIV DNA Integration." Cold Spring Harbor Perspectives in Medicine 2(7).

Craigo, J. K. and P. Gupta (2006). "HIV-1 in Genital Compartments: Vexing Viral Reservoirs." Current Opinion in HIV and AIDS 1(2):97.

D'Aquila, R. T., M. D. Hughes, et al. (1996). "Nevirapine, Zidovudine, and Didanosine Compared with Zidovudine and Didanosine in Patients with HIV-1 Infection. A Randomized, Double-Blind, Placebo-Controlled Trial. National Institute of Allergy and Infectious Diseases AIDS Clinical Trials Group Protocol 241 Investigators." Annals of Internal Medicine 124(12):1019.

de Béthune, M.-P. (2010). "Non-nucleoside Reverse Transcriptase Inhibitors (NNRTIs), Their Discovery, Development, and Use in the Treatment of HIV-1 Infection: A Review of the Last 20 Years (1989–2009)." Antiviral Research 85(1):75.

De Cock, K. M., G. Adjorlolo, et al. (1993). "Epidemiology and Transmission of HIV-2: Why There is No HIV-2 Pandemic." JAMA 270(17):2083.

De Oliveira, T., K. Deforche, et al. (2005). "An Automated Genotyping System for Analysis of HIV-1 and Other Microbial Sequences." Bioinformatics 21(19):3797.

Delgado, E., W. K. Ampofo, et al. (2008). "High Prevalence of Unique Recombinant Forms of HIV-1 in Ghana: Molecular Epidemiology from an Antiretroviral Resistance Study." Journal of Acquired Immune Deficiency Syndromes 48(5):599.

Domingo, E. (2002). "Quasispecies Theory in Virology." J Virol 76(1):463.

Domingo, E., J. Sheldon, et al. (2012). "Viral Quasispecies Evolution." Microbiology and Molecular Biology Reviews 76(2):159.

Donnell, D., J. M. Baeten, et al. (2010). "Heterosexual HIV-1 Transmission After Initiation of Antiretroviral Therapy: A Prospective Cohort Analysis." The Lancet 375(9731):2092.

Dudley, D. M., E. N. Chin, et al. (2012). "Low-cost Ultra-wide Genotyping Using Roche/454 Pyrosequencing for Surveillance of HIV Drug Resistance." PloS One 7(5):e36494.

Eames, K. T. D. and M. J. Keeling (2003). "Contact Tracing and Disease Control." Proceedings of the Royal Society of London. Series B: Biological Sciences 270(1533):2565.

Edgar, R. C. (2004). "MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput." Nucleic Acids Research 32(5):1792.

Edgar, R. C. (2013). "UPARSE: Highly Accurate OTU Sequences from Microbial Amplicon Reads." Nature Methods 10(10):996.

Eigen, M. (1971). "Selforganization of Matter and the Evolution of Biological Macromolecules." Naturwissenschaften 58(10):465.

Embretson, J., M. Zupancic, et al. (1993). "Massive Covert Infection of Helper T Lymphocytes and Macrophages by HIV During the Incubation Period of AIDS." Nature 362(6418):359.

Engelman, A. and P. Cherepanov (2012). "The Structural Biology of HIV-1: Mechanistic and Therapeutic Insights." Nature Reviews Microbiology 10(4):279.

Farnet, C. M. and W. A. Haseltine (1991). "Determination of Viral Proteins Present in the Human Immunodeficiency Virus Type 1 Pre-integration Complex." J Virol 65(4):1910.

Fassati, A. (2012). "Multiple Roles of the Capsid Protein in the Early Steps of HIV-1 Infection." Virus Research 170:15.

Fauci, A. S. (1993). "HIV Infection is Active and Progressive in Lymphoid Ttissue During the Clinically Latent Stage of Disease." Nature 362:25.

Felsenstein, J. (1985). "Confidence Limits on Phylogenies: An Approach Using the Bootstrap." Evolution:783.

Fiebig, E. W., D. J. Wright, et al. (2003). "Dynamics of HIV Viremia and Antibody Seroconversion in Plasma Donors: Implications for Diagnosis and Staging of Primary HIV Infection." AIDS 17(13):1871.

Finzi, D., S. F. Plaeger, et al. (2006). "Defective Virus Drives Human Immunodeficiency Virus Infection, Persistence, and Pathogenesis." Clinical and Vaccine Immunology 13(7):715.

Fischer, W., V. V. Ganusov, et al. (2010). "Transmission of Single HIV-1 Genomes and Dynamics of Early Immune Escape Revealed by Ultra-deep Sequencing." PloS One 5(8): e12303.

Fischetti, L., O. Opare-Sem, et al. (2004). "Higher Viral Load May Explain the Dominance of CRF02_AG in the Molecular Epidemiology of HIV in Ghana." AIDS 18(8):1208.

Fischetti, L., O. Opare-Sem, et al. (2004). "Molecular Epidemiology of HIV in Ghana: Dominance of CRF02_AG." Journal of Medical Virology 73(2):158.

Fischl, M. A., D. D. Richman, et al. (1987). "The Efficacy of Azidothymidine (AZT) in the Treatment of Patients with AIDS and AIDS-related Complex." New England Journal of Medicine 317(4):185.

Fiscus, S. A., C. D. Pilcher, et al. (2007). "Rapid, Real-time Detection of Acute HIV Infection in Patients in Africa." Journal of Infectious Diseases 195(3):416.

Fisher, M., D. Pao, et al. (2010). "Determinants of HIV-1 Transmission in Men who have Sex with Men: A Combined Clinical, Epidemiological and Phylogenetic Approach." AIDS 24(11):1739.

Forshey, B. M., U. von Schwedler, et al. (2002). "Formation of a Human Immunodeficiency Virus type 1 Core of Optimal Stability is Crucial for Viral Replication." J Virol 76(11):5667.

Foster, G. M., J. C. Ambrose, et al. (2014). "Novel HIV-1 Recombinants Spreading across Multiple Risk Groups in the United Kingdom: The Identification and Phylogeography of Circulating Recombinant Form (CRF) 50_A1D." PloS One 9(1):e83337.

Fox, J., H. Castro, et al. (2010). "Epidemiology of Non-B clade forms of HIV-1 in Men who have Sex with Men in the UK." AIDS 24(15):2397.

Frankel, A. D. and J. A. T. Young (1998). "HIV-1: Fifteen Proteins and an RNA." Annual Review of Biochemistry 67(1):1.

Frentz, D., C. A. Boucher, et al. (2012). "Temporal Changes in the Epidemiology of Transmission of Drug-resistant HIV-1 Across the World." AIDS Rev 14(1):17.

Friedman-Kien, A. E., L. Laubenstein, et al. (1981). "Kaposis Sarcoma and Pneumocystis Pneumonia Among Homosexual Men--New York City and California." Morbidity and Mortality Weekly Report 30(25):305.

García-Lerma, J. G., S. Nidtha, et al. (2001). "Increased Ability for Selection of Zidovudine Resistance in a Distinct Class of Wild-Type HIV-1 from Drug-Naive Persons." Proceedings of the National Academy of Sciences 98(24):13907.

Garrett, N. J., S. Lattimore, et al. (2012). "The Recent Infection Testing Algorithm (RITA) in Clinical Practice: A Survey of HIV Clinicians in England and Northern Ireland." HIV Medicine 13(7):444.

Gasper-Smith, N., D. M. Crossman, et al. (2008). "Induction of Plasma (TRAIL), TNFR-2, Fas Ligand, and Plasma Microparticles After Human Immunodeficiency Virus type 1 (HIV-1) Transmission: Implications for HIV-1 Vaccine Design." J Virol 82(15):7700.

Geretti, A. M. (2006). "HIV-1 Subtypes: Epidemiology and Significance for HIV Management." Current Opinion in Infectious Diseases 19(1):1.

Geretti, A. M., L. Harrison, et al. (2009). "Effect of HIV-1 Subtype on Virologic and Immunologic Response to Starting Highly Active Antiretroviral Therapy." Clinical Infectious Diseases 48(9):1296.

Gianella, S., W. Delport, et al. (2011). "Detection of Minority Resistance During Early HIV-1 Infection: Natural Variation and Spurious Detection rather than Transmission and Evolution of Multiple Viral Variants." J Virol 85(16):8359.

Gifford, R. J., T. F. Liu, et al. (2009). "The Calibrated Population Resistance Tool: Standardized Genotypic Estimation of Transmitted HIV-1 Drug Resistance." Bioinformatics 25(9):1197.

Giorgi, E., B. Funkhouser, et al. (2010). "Estimating Time Since Infection in Early Homogeneous HIV-1 Samples using a Poisson Model." BMC Bioinformatics 11(1):532.

Girardi, E., C. A. Sabin, et al. (2007). "Late Diagnosis of HIV Infection: Epidemiological Features, Consequences and Strategies to Encourage Earlier Testing." Journal of Acquired Immune Deficiency Syndromes 46:S3.

Glenn, T. C. (2011). "Field Guide to Next-generation DNA Sequencers." Molecular Ecology Resources 11(5):759.

Gottlieb, M. S., R. Schroff, et al. (1981). "Pneumocystis Carinii Pneumonia and Mucosal Candidiasis in Previously Healthy Homosexual Men: Evidence of a New Acquired Cellular Immunodeficiency." The New England Journal of Medicine 305(24):1425.

Granich, R. M., C. F. Gilks, et al. (2009). "Universal Voluntary HIV Testing with Immediate Antiretroviral Therapy as a Strategy for Elimination of HIV Transmission: A Mathematical Model." The Lancet 373(9657):48.

Grant, R. M., D. R. Kuritzkes, et al. (2003). "Accuracy of the TRUGENE HIV-1 Genotyping Kit." Journal of Clinical Microbiology 41(4):1586.

Gray, L., J. Sterjovski, et al. (2005). "Uncoupling Coreceptor Usage of Human Immunodeficiency Virus Type 1 (HIV-1) from Macrophage Tropism Reveals Biological Properties of CCR5-restricted HIV-1 Isolates from Patients with Acquired Immunodeficiency Syndrome." Virology 337(2):384.

Guindon, S., J.-F. Dufayard, et al. (2010). "New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0." Systematic Biology 59(3):307.

Gulick, R. M., J. W. Mellors, et al. (1997). "Treatment with Indinavir, Zidovudine, and Lamivudine in Adults with Human Immunodeficiency Virus Infection and Prior Antiretroviral Therapy." New England Journal of Medicine 337(11):734.

Gupta, R. K., M. R. Jordan, et al. (2012). "Global Trends in Antiretroviral Resistance in Treatment-Naive Individuals with HIV After Rollout of Antiretroviral Treatment in Resource-Limited Settings: A Global Collaborative Study and Meta-Regression Analysis." The Lancet 9849:1250.

Haase, A. T. (2010). "Targeting Early Infection to Prevent HIV-1 Mucosal Transmission." Nature 464(7286):217.

Haggerty, S. and M. Stevenson (1991). "Predominance of Distinct Viral Genotypes in Brain and Lymph Node Compartments of HIV-1-infected Individuals." Viral Immunology 4(2):123.

Hamers, R. L., C. L. Wallis, et al. (2011). "HIV-1 Drug Resistance in Antiretroviral-Naive Individuals in Sub-Saharan Africa After Rollout of Antiretroviral Therapy: A Multicentre Observational Study." The Lancet Infectious Diseases 11(10):750.

Hammer, S. M., D. A. Katzenstein, et al. (1996). "A Trial Comparing Nucleoside Monotherapy with Combination Therapy in HIV-Infected Adults with CD4 Cell Counts from 200 to 500 per Cubic Millimeter." New England Journal of Medicine 335(15):1081.

Hammer, S. M., K. E. Squires, et al. (1997). "A Controlled Trial of Two Nucleoside Analogues Plus Indinavirin Persons with Human Immunodeficiency Virus Infection and CD4 Cell Counts of 200 per Cubic Millimeter or Less." New England Journal of Medicine 337(11):725.

Harris, C., C. B. Small, et al. (1983). "Immunodeficiency in Female Sexual Partners of Men with the Acquired Immunodeficiency Syndrome." The New England Journal of Medicine 308(20):1181.

Hawkins, D. M., J. A. Garrett, et al. (2001). "Some Issues in Resolution of Diagnostic Tests Using an Imperfect Gold Standard." Statistics in Medicine 20(13):1987.

Health Protection Agency (2012). HIV in the United Kingdom: 2012 Report. London: Health Protection Services, Colindale, Health Protection Agency.

Hedskog, C., M. Mild, et al. (2010). "Dynamics of HIV-1 Quasispecies During Antiviral Treatment Dissected using Ultra-deep Pyrosequencing." PloS One 5(7):e11345.

Hemelaar, J., E. Gouws, et al. (2011). "Global Trends in Molecular Epidemiology of HIV-1 During 2000–2007." AIDS 25(5):679.

Hemelaar, J., E. Gouws, et al. (2011). "Global Trends in Molecular Epidemiology of HIV-1 During 2000–2007." AIDS 25(5):679.

Henn, M. R., C. L. Boutwell, et al. (2012). "Whole Genome Deep sequencing of HIV-1 Reveals the Impact of Early Minor Variants Upon Immune Recognition During Acute Infection." PLoS Pathogens 8(3):e1002529.

Hillis, D. M. and J. P. Huelsenbeck (1994). "Support for Dental HIV Transmission." Nature 369:24.

Hladik, F. and T. J. Hope (2009). "HIV Infection of the Genital Mucosa in Women." Current HIV/AIDS Reports 6(1):20.

Hladik, F. and M. J. McElrath (2008). "Setting the Stage: Host Invasion by HIV." Nature Reviews Immunology 8(6):447.

Ho, S. (2008). "The Molecular Clock and Estimating Species Divergence." Nature Education 1(1):1.

Hogben, M., T. McNally, et al. (2007). "The Effectiveness of HIV Partner Counseling and Referral Services in Increasing Identification of HIV-Positive Individuals: A Systematic Review." American Journal of Preventive Medicine 33(2):S89.

Hollingsworth, T. D., R. M. Anderson, et al. (2008). "HIV-1 Transmission, by Stage of Infection." Journal of Infectious Diseases 198(5):687.

House, T. and M. J. Keeling (2010). "The Impact of Contact Tracing in Clustered Populations." PLoS Computational Biology 6(3):e1000721.

Huang, W., A. Gamarnik, et al. (2003). "Amino Acid Substitutions at Position 190 of Human Immunodeficiency Virus Type 1 Reverse Transcriptase Increase Susceptibility to Delavirdine and Impair Virus Replication." J Virol 77(2):1512.

Huang, X. and J. Zhang (1996). "Methods for Comparing a DNA Sequence with a Protein Sequence." Computer applications in the biosciences: CABIOS 12(6):497.

Hué, S., J. P. Clewley, et al. (2004). "HIV-1 *pol* Gene Variation is Sufficient for Reconstruction of Transmissions in the Era of Antiretroviral Therapy." AIDS 18(5):719.

Hué, S., J. P. Clewley, et al. (2005). "Investigation of HIV-1 Transmission Events by Phylogenetic Methods: Requirement for Scientific Rigour." AIDS 19(4):449.

Hué, S., R. J. Gifford, et al. (2009). "Demonstration of Sustained Drug-resistant Human Immunodeficiency Virus Type 1 Lineages Circulating Among Treatment-naive Individuals." J Virol 83(6):2645.

Huelsenbeck, J. P., B. Larget, et al. (2000). "A Compound Poisson Process for Relaxing the Molecular Clock." Genetics 154(4):1879.

Hughes, J. P. and P. Totten (2003). "Estimating the Accuracy of Polymerase Chain Reaction–Based Tests Using Endpoint Dilution." Biometrics 59(3):505.

Hulme, A. E., O. Perez, et al. (2011). "Complementary Assays Reveal a Relationship Between HIV-1 Uncoating and Reverse Transcription." Proceedings of the National Academy of Sciences 108(24):9975.

Huse, S. M., J. A. Huber, et al. (2007). "Accuracy and Quality of Massively Parallel DNA Pyrosequencing." Genome Biol 8(7):R143.

Ishikawa, K., W. Janssens, et al. (1996). "Sequence Note: Genetic and Phylogenetic Analysis of HIV Type 1 *env* Subtypes in Ghana, West Africa." AIDS Research and Human Retroviruses 12(16):1575.

Jabara, C. B., C. D. Jones, et al. (2011). "Accurate Sampling and Deep Sequencing of the HIV-1 Protease Gene using a Primer ID." Proceedings of the National Academy of Sciences 108(50):20166.

Jacquez, J. A., J. S. Koopman, et al. (1994). "Role of the Primary Infection in Epidemics of HIV Infection in Gay Cohorts." Journal of Acquired Immune Deficiency Syndromes 7(11):1169.

Janssen, R. S., G. A. Satten, et al. (1998). "New Testing Strategy to Detect Early HIV-1 Infection for Use in Incidence Estimates and for Clinical and Prevention Purposes." JAMA 280(1):42.

Ji, H., Y. Li, et al. (2011). "Next-generation Sequencing of Dried Blood Spot Specimens: A Novel Approach to HIV Drug-resistance Surveillance." Antiviral Therapy 16(6):871.

Jiang, J., S. D. Ablan, et al. (2011). "The Interdomain Linker Region of HIV-1 Capsid Protein is a Critical Determinant of Proper Core Assembly and Stability." Virology 421(2):253.

Jolly, C., N. J. Booth, et al. (2010). "Cell-cell Spread of Human Immunodeficiency Virus Type 1 Overcomes Tetherin/BST-2-Mediated Restriction in T Cells." J Virol 84(23):12185.

Jolly, C. and Q. J. Sattentau (2004). "Retroviral Spread by Induction of Virological Synapses." Traffic 5(9):643.

Jones, K. A. and M. B. Peterlin (1994). "Control of RNA Initiation and Elongation at the HIV-1 Promoter." Annual Review of Biochemistry 63(1):717.

Kafaie, J., R. Song, et al. (2008). "Mapping of Nucleocapsid Residues Important for HIV-1 Genomic RNA Dimerization and Packaging." Virology 375(2):592.

Kanagawa, T. (2003). "Bias and Artifacts in Multitemplate Polymerase Chain Reactions (PCR)." Journal of Bioscience and Bioengineering 96(4):317.

Karita, E., M. Price, et al. (2007). "Investigating the Utility of the HIV-1 BED Capture Enzyme Immunoassay Using Cross-sectional and Longitudinal Seroconverter Specimens from Africa." AIDS 21(4):403.

Kaye, M., D. Chibo, et al. (2008). "Phylogenetic Investigation of Transmission Pathways of Drug-resistant HIV-1 Utilizing *pol* Sequences Derived from Resistance Genotyping." Journal of Acquired Immune Deficiency Syndromes 49(1):9.

Keele, B. F. and J. D. Estes (2011). "Barriers to Mucosal Transmission of Immunodeficiency Viruses." Blood 118(4):839.

Keele, B. F., E. E. Giorgi, et al. (2008). "Identification and Characterization of Transmitted and Early Founder Virus Envelopes in Primary HIV-1 Infection." Proceedings of the National Academy of Sciences 105(21):7552.

Keele, B. F., H. Li, et al. (2009). "Low-dose Rectal Inoculation of Rhesus Macaques by SIVsmE660 or SIVmac251 Recapitulates Human Mucosal Infection by HIV-1." The Journal of Experimental Medicine 206(5):1117.

Kelley, C. F., J. D. Barbour, et al. (2007). "The Relation Between Symptoms, Viral Load, and Viral Load Set Point in Primary HIV Infection." Journal of Acquired Immune Deficiency Syndromes 45(4):445.

Kilby, J. M., S. Hopkins, et al. (1998). "Potent Suppression of HIV-1 Replication in Humans by T-20, a Peptide Inhibitor of gp41-mediated Virus Entry." Nature Medicine 4(11):1302.

Kim, A. A., T. Hallett, et al. (2011). "Estimating HIV Incidence Among Adults in Kenya and Uganda: A Systematic Comparison of Multiple Methods." PloS One 6(3):e17535.

Kim, A. A., J. S. McDougal, et al. (2010). "Evaluating the BED Capture Enzyme Immunoassay to Estimate HIV Incidence Among Adults in Three Countries in sub-Saharan Africa." AIDS Research and Human Retroviruses 26(10):1051.

Kiss, I. Z., D. M. Green, et al. (2005). "Disease Contact Tracing in Random and Clustered Networks." Proceedings of the Royal Society B: Biological Sciences 272(1570):1407.

Klein, J. S. and P. J. Bjorkman (2010). "Few and Far Between: How HIV May Be Evading Antibody Avidity." PLoS Pathogens 6(5):e1000908.

Koboldt, D. C., K. M. Steinberg, et al. (2013). "The Next-generation Sequencing Revolution and its Impact on Genomics." Cell 155(1):27.

Kondo, E., F. Mammano, et al. (1995). "The p6-gag Domain of Human Immunodeficiency Virus type 1 is Sufficient for the Incorporation of Vpr into Heterologous Viral Particles." J Virol 69(5):2759.

Konings, F. A. J., S. T. Burda, et al. (2006). "Human Immunodeficiency Virus type 1 (HIV-1) Circulating Recombinant Form 02_AG (CRF02_AG) has a Higher In Vitro Replicative Capacity than its Parental Subtypes A and G." Journal of Medical Virology 78(5):523.

Koopman, J. S., J. A. Jacquez, et al. (1997). "The Role of Early HIV Infection in the Spread of HIV Through Populations." Journal of Acquired Immune Deficiency Syndromes 14(3):249.

Korber, B., M. Muldoon, et al. (2000). "Timing the Ancestor of the HIV-1 Pandemic Strains." Science 288(5472):1789.

Kouyos, R. D., V. von Wyl, et al. (2011). "Ambiguous Nucleotide Calls from Population-based Sequencing of HIV-1 Are a Marker For Viral Diversity and the Age of Infection." Clinical Infectious Diseases 52(4):532.

Lackner, A. A. and R. S. Veazey (2007). "Current Concepts in AIDS Pathogenesis: Insights from the SIV/macaque Model." Annu. Rev. Med. 58:461.

Laird, G. M., E. E. Eisele, et al. (2013). "Rapid Quantification of the Latent Reservoir for HIV-1 Using a Viral Outgrowth Assay." PLoS Pathogens 9(5):e1003398.

Lander, E. S., L. M. Linton, et al. (2001). "Initial Sequencing and Analysis of the Human Genome." Nature 409(6822):860.

Lauring, A. S. and R. Andino (2010). "Quasispecies Theory and the Behavior of RNA Viruses." PLoS Pathogens 6(7):e1001005.

Lavreys, L., J. M. Baeten, et al. (2002). "Virus Load During Primary Human Immunodeficiency Virus (HIV) type 1 Infection is Related to the Severity of Acute HIV Illness in Kenyan Women." Clinical Infectious Diseases 35(1):77.

Leitner, T. and J. Albert (1999). "The Molecular Clock of HIV-1 Unveiled Through Analysis of a Known Transmission History." Proceedings of the National Academy of Sciences 96(19):10752.

Levy, J. A., A. D. Hoffman, et al. (1984). "Isolation of Lymphocytopathic Retroviruses from San Francisco Patients with AIDS." Science 225(4664):840.

Lewis, F., G. J. Hughes, et al. (2008). "Episodic Sexual Transmission of HIV Revealed by Molecular Phylodynamics." PLoS Medicine 5(3):e50.

Leynaert, B., A. M. Downs, et al. (1998). "Heterosexual Transmission of Human Immunodeficiency Virus Variability of Infectivity throughout the Course of Infection." American Journal of Epidemiology 148(1):88.

Li, H., K. J. Bar, et al. (2010). "High Multiplicity Infection by HIV-1 in Men who have Sex with Men." PLoS pathogens 6(5):e1000890.

Li, J. Z., R. Paredes, et al. (2013). "Impact of Minority Nonnucleoside Reverse Transcriptase Inhibitor Resistance Mutations on Resistance Genotype After Virologic Failure." Journal of Infectious Diseases 207(6):893.

Li, X., C. Ning, et al. (2013). "Genome Sequences of a Novel HIV-1 Circulating Recombinant Form (CRF61_BC) Identified Among Heterosexuals in China." Genome Announcements 1(3):e00326.

Locatelli, S. and M. Peeters (2012). "Cross-species Transmission of Simian Retroviruses: How and Why They Could Lead to the Emergence of New Diseases in the Human Population." AIDS 26(6):659.

Lodwick, R., A. Alioum, et al. (2011). "HIV in Hiding: Methods and Data Requirements for the Estimation of the Number of People Living with Undiagnosed HIV, Working Group on Estimation of HIV Prevalence in Europe." AIDS 25(8):1017.

Lohr, S. (2009). Sampling: Design and Analysis, Cengage Learning.

Lynch, R. M., T. Shen, et al. (2009). "Appreciating HIV type 1 Diversity: Subtype Differences in Env." AIDS Research and Human Retroviruses 25(3):237.

Malim, M. H., J. Hauber, et al. (1989). "The HIV-1 Rev Trans-activator Acts Through a Structured Target Sequence to Activate Nuclear Export of Unspliced Viral mRNA." Nature 338(6212):254.

Manavi, K., A. McMillan, et al. (2004). "Heterosexual Men and Women with HIV Test Positive at a Later Stage of Infection Than Homo-or Bisexual Men." International Journal of STD & AIDS 15(12):811.

Mann, D. A., I. Mikaélian, et al. (1994). "A Molecular Rheostat: Co-operative Rev Binding to Stem I of the Rev-response Element Modulates Human Immunodeficiency Virus type-1 Late Gene Expression." Journal of Molecular Biology 241(2):193.

Marks, G., N. Crepaz, et al. (2006). "Estimating Sexual Transmission of HIV from Persons Aware and Unaware That They Are Infected with the Virus in the USA." AIDS 20(10):1447.

Marks, G., N. Crepaz, et al. (2005). "Meta-analysis of High-risk Sexual Behavior in Persons Aware and Unaware They Are Infected with HIV in the United States: Implications for HIV Prevention Programs." Journal of Acquired Immune Deficiency Syndromes 39(4):446.

Martin, M. P., X. Gao, et al. (2002). "Epistatic Interaction Between KIR3DS1 and HLA-B Delays the Progression to AIDS." Nature Genetics 31(4):429.

Martin, M. P., Y. Qi, et al. (2007). "Innate Partnership of HLA-B and KIR3DL1 Subtypes Against HIV-1." Nature Genetics 39(6):733.

Marx, P. A., P. G. Alcabes, et al. (2001). "Serial Human Passage of Simian Immunodeficiency Virus by Unsterile Injections and the Emergence of Epidemic Human Immunodeficiency Virus in Africa." Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 356(1410):911.

Mastro, T. D., A. A. Kim, et al. (2010). "Estimating HIV Incidence in Populations Using Tests for Recent Infection: Issues, Challenges and the Way Forward." Journal of HIV AIDS Surveillance & Epidemiology 2(1):1.

Masur, H., M. A. Michelis, et al. (1982). "Opportunistic Infection in Previously Healthy Women Initial Manifestations of a Community-Acquired Cellular Immunodeficiency." Annals of Internal Medicine 97(4):533.

Mathers, B. M., L. Degenhardt, et al. (2008). "Global Epidemiology of Injecting Drug Use and HIV Among People Who Inject Drugs: A Systematic Review." The Lancet 372(9651):1733.

McMichael, A. J., P. Borrow, et al. (2009). "The Immune Response During Acute HIV-1 Infection: Clues for Vaccine Development." Nature Reviews Immunology 10(1):11.

Melikyan, G. B., R. M. Markosyan, et al. (2000). "Evidence That the Transition of HIV-1 gp41 into a Six-helix Bundle, Not the Bundle Configuration, Induces Membrane Fusion." The Journal of Cell Biology 151(2):413.

Metzker, M. L. (2009). "Sequencing Technologies—the Next Generation." Nature Reviews Genetics 11(1):31.

Miller, C. J., N. J. Alexander, et al. (1989). "Genital Mucosal Transmission of Simian Immunodeficiency Virus: Animal Model for Heterosexual Transmission of Human Immunodeficiency Virus." J Virol 63(10):4277.

Miller, C. J., Q. Li, et al. (2005). "Propagation and Dissemination of Infection After Vaginal Transmission of Simian Immunodeficiency Virus." J Virol 79(14):9217.

Miller, W. C., N. E. Rosenberg, et al. (2010). "The Role of Acute and Early HIV Infection in the Sexual Transmission of HIV." Current Opinion in HIV and AIDS 5(4):277.

Mitsuya, H., K. J. Weinhold, et al. (1985). "3'-Azido-3'-deoxythymidine (BW A509U): An Antiviral Agent That Inhibits the Infectivity and Cytopathic Effect of Human T-lymphotropic Virus Type III/lymphadenopathy-associated Virus In Vitro." Proceedings of the National Academy of Sciences 82(20):7096.

Mogensen, T. H., J. Melchjorsen, et al. (2010). "Review Innate Immune Recognition and Activation During HIV Infection." Retrovirology 7(1):54

Morrison, C. S., K. Demers, et al. (2010). "Plasma and Cervical Viral Loads Among Ugandan and Zimbabwean Women During Acute and Early HIV-1 Infection." AIDS 24(4):573.

Moya, A., E. C. Holmes, et al. (2004). "The Population Genetics and Evolutionary Epidemiology of RNA Viruses." Nature Reviews Microbiology 2(4):279.

NAM (2013). "http://www.aidsmap.com/resources/Antiretroviral-drugs-chart/page/1412453/."

National HIV/AIDS/STI Control Programme (2010). "Guidelines for Antiretroviral Therapy in Ghana."

Nazli, A., O. Chan, et al. (2010). "Exposure to HIV-1 Directly Impairs Mucosal Epithelial Barrier Integrity Allowing Microbial Translocation." PLoS Pathogens 6(4):e1000852.

Newman, M. E. J. (2003). "Mixing Patterns in Networks." Physical Review E 67(2):026126.

Nii-Trebi, N. I., S. Ibe, et al. (2013). "HIV-1 Drug-Resistance Surveillance Among Treatment-Experienced and-Naïve Patients After the Implementation of Antiretroviral Therapy in Ghana." PloS One 8(8):e71972.

Njai, H. F., Y. Gali, et al. (2006). "The Predominance of Human Immunodeficiency Virus type 1 (HIV-1) Circulating Recombinant Form 02 (CRF02_AG) in West Central Africa May Be Related to its Replicative Fitness." Retrovirology 3(1):40.

Nowak, M. A., S. Bonhoeffer, et al. (1997). "Anti-viral Drug Treatment: Dynamics of Resistance in Free Virus and Infected Cell Populations." Journal of Theoretical Biology 184(2):203.

Ohene, S. and E. Forson (2009). "Care of Patients on Anti-Retroviral Therapy in Kumasi Metropolis." Ghana Medical Journal 43(4):144.

Oksanen, J. F., G. Blanchet, et al. (2013). "Vegan: Community Ecology Package."

Onafuwa-Nuga, A. and A. Telesnitsky (2009). "The Remarkable Frequency of Human Immunodeficiency Virus type 1 Genetic Recombination." Microbiology and Molecular Biology Reviews 73(3):451.

Ott, M., M. Geyer, et al. (2011). "The Control of HIV Transcription: Keeping RNA Polymerase II on Track." Cell Host & Microbe 10(5):426.

Palmer, S., A. P. Wiegand, et al. (2003). "New Real-Time Reverse Transcriptase-Initiated PCR Assay with Single-Copy Sensitivity for Human Immunodeficiency Virus Type 1 RNA in Plasma." Journal of Clinical Microbiology 41(10):4531.

Pao, D., M. Fisher, et al. (2005). "Transmission of HIV-1 During Primary Infection: Relationship to Sexual Risk and Sexually Transmitted Infections." AIDS 19(1):85.

Paradis, E., J. Claude, et al. (2004). "APE: Analyses of Phylogenetics and Evolution in R Language." Bioinformatics 20(2):289.

Paraskevis, D., E. Magiorkinis, et al. (2004). "Phylogenetic Reconstruction of a Known HIV-1 CRF04_cpx Transmission Network Using Maximum Likelihood and Bayesian Methods." Journal of Molecular Evolution 59(5):709.

Parekh, B. S., M. S. Kennedy, et al. (2002). "Quantitative Detection of Increasing HIV Type 1 Antibodies After Seroconversion: a Simple Assay for Detecting Recent HIV Infection and Estimating Incidence." AIDS Research and Human Retroviruses 18(4):295.

Parrish, N. F., F. Gao, et al. (2013). "Phenotypic Properties of Transmitted Founder HIV-1." Proceedings of the National Academy of Sciences 110(17):6626.

Patient, D. (1992). "Molecular Epidemiology of HIV Transmission in a Dental Practice." Science 256:22.

Patton, S. J., A. J. Wallace, et al. (2006). "Benchmark for Evaluating the Quality of DNA Sequencing: Proposal from an International External Quality Assessment Scheme." Clinical Chemistry 52(4):728.

Peeters, M., V. Courgnaud, et al. (2002). "Risk to Human Health from a Plethora of Simian Immunodeficiency Viruses in Primate Bushmeat." Emerging Infectious Diseases 8(5):451.

Peeters, M., C. Toure-Kane, et al. (2003). "Genetic Diversity of HIV in Africa: Impact on Diagnosis, Treatment, Vaccine Development and Trials." AIDS 17(18):2547.

Pereira, L. A., K. Bentley, et al. (2000). "Survey and Summary: a Compilation of Cellular Transcription Factor Interactions with the HIV-1 LTR Promoter." Nucleic Acids Research 28(3):663.

Perelson, A. S. (2002). "Modelling Viral and Immune System Dynamics." Nature Reviews Immunology 2(1):28.

Piguet, V. and Q. Sattentau (2004). "Dangerous Liaisons at the Virological Synapse." Journal of Clinical Investigation 114(5):605.

Pilcher, C. D., G. Joaki, et al. (2007). "Amplified Transmission of HIV-1: Comparison of HIV-1 Concentrations in Semen and Blood During Acute and Chronic Infection." AIDS 21(13):1723.

Pilcher, C. D., D. C. Shugars, et al. (2001). "HIV in Body Fluids During Primary HIV Infection: Implications for Pathogenesis, Treatment and Public Health." AIDS 15(7):837.

Pilcher, C. D., H. C. Tien, et al. (2004). "Brief But Efficient: Acute HIV Infection and the Sexual Transmission of HIV." Journal of Infectious Diseases 189(10):1785.

Pineda-Peña, A.-C., N. R. Faria, et al. (2013). "Automated Subtyping of HIV-1 Genetic Sequences for Clinical and Surveillance." Infect Genet Evol. 2013 Oct;19:337

Pinkerton, S. D. (2007). "How Many Sexually-Acquired HIV Infections in the USA Are Due to Acute-Phase HIV Transmission?" AIDS, 21(12):1625.

Pitchenik, A. E., M. A. Fischl, et al. (1983). "Opportunistic Infections and Kaposi's Sarcoma Among Haitians: Evidence of a New Acquired Immunodeficiency State." Annals of Internal Medicine 98(3):277.

Pommier, Y., A. A. Johnson, et al. (2005). "Integrase Inhibitors to Treat HIV/AIDS." Nature Reviews Drug Discovery 4(3):236.

Poon, A. F. Y., R. A. McGovern, et al. (2011). "Dates of HIV Infection Can Be Estimated for Seroprevalent Patients by Coalescent Analysis of Serial Next-Generation Sequencing Data." AIDS 25(16):2019.

Price, M. N., P. S. Dehal, et al. (2010). "FastTree 2–Approximately Maximum-Likelihood Trees for Large Alignments." PloS One 5(3):e9490.

Ragonnet-Cronin, M., S. Aris-Brosou, et al. (2012). "Genetic Diversity as a Marker for Timing Infection in HIV-Infected Patients: Evaluation of a 6-Month Window and Comparison with BED." Journal of Infectious Diseases 206(5):756.

Rambaut, A., D. Posada, et al. (2004). "The Causes and Consequences of HIV Evolution." Nature Reviews Genetics 5(1):52.

Rawal, B. D., A. Degula, et al. (2003). "Development of a New Less-Sensitive Enzyme Immunoassay for Detection of Early HIV-1 Infection." Journal of Acquired Immune Deficiency Syndromes 33(3):349.

Redd, A. D., C. E. Mullis, et al. (2012). "The Rates of HIV Superinfection and Primary HIV Incidence in a General Population in Rakai, Uganda." Journal of Infectious Diseases 206(2):267.

Redd, A. D., T. C. Quinn, et al. (2013). "Frequency and Implications of HIV Superinfection." The Lancet Infectious Diseases 13(7):622.

Robertson, D. L., J. P. Anderson, et al. (1999). "HIV-1 Nomenclature Proposal." Human Retroviruses and AIDS:492.

Robertson, D. L., P. M. Sharp, et al. (1995). "Recombination in HIV-1." Nature 374(6518):124.

Robertson, J. R., A. B. Bucknall, et al. (1986). "Epidemic of AIDS Related Virus (HTLV-III/LAV) Infection Among Intravenous Drug Abusers." British Medical Journal 292(6519):527.

Rozenbaum, W., J. Coulaud, et al. (1982). "Multiple Opportunistic Infection in a Male Homosexual in France." The Lancet 319(8271):572.

Rozera, G., I. Abbate, et al. (2009). "Massively Parallel Pyrosequencing Highlights Minority Variants in the HIV-1 Env Quasispecies Deriving from Lymphomonocyte Sub-Populations." Retrovirology 6(1):15.

Rutjes, A. W. S., J. B. Reitsma, et al. (2007). "Evaluation of Diagnostic Tests When There is No Gold Standard: A Review of Methods" Health Technol Assess. 2007 Dec;11(50):iii.

Sagoe, K. W., M. Dwidar, et al. (2009). "HIV-1 CRF 02 AG Polymerase Genes in Southern Ghana are Mosaics of Different 02 AG Strains and the Protease Gene Cannot Infer Subtypes." Virology Journal 6(1):27.

Sagoe, K. W. C., M. Dwidar, et al. (2007). "Variability of the Human Immunodeficiency Virus Type 1 Polymerase Gene from Treatment Naive Patients in Accra, Ghana." Journal of Clinical Virology 40(2):163.

Sakarovitch, C., F. Rouet, et al. (2007). "Do Tests Devised to Detect Recent HIV-1 Infection Provide Reliable Estimates of Incidence in Africa?" Journal of Acquired Immune Deficiency Syndromes 45(1):115.

Salazar-Gonzalez, J. F., M. G. Salazar, et al. (2009). "Genetic Identity, Biological Phenotype, and Evolutionary Pathways of Transmitted/Founder Viruses in Acute and Early HIV-1 Infection." The Journal of Experimental Medicine 206(6):1273.

Sanger, F., S. Nicklen, et al. (1977). "DNA Sequencing with Chain-Terminating Inhibitors." Proceedings of the National Academy of Sciences 74(12):5463.

Scaduto, D. I., J. M. Brown, et al. (2010). "Source Identification in Two Criminal Cases Using Phylogenetic Analysis of HIV-1 DNA Sequences." Proceedings of the National Academy of Sciences 107(50):21242.

Schuitemaker, H., M. Koot, et al. (1992). "Biological Phenotype of Human Immunodeficiency Virus type 1 Clones at Different Stages of Infection: Progression of Disease is Associated with a Shift from Monocytotropic to T-Cell-Tropic Virus Population." J Virol 66(3):1354.

Semaille, C., F. Barin, et al. (2007). "Monitoring the Dynamics of the HIV Epidemic Using Assays for Recent Infection and Serotyping Among New HIV Diagnoses: Experience After 2 Years in France." Journal of Infectious Diseases 196(3):377.

Shankarappa, R., J. B. Margolick, et al. (1999). "Consistent Viral Evolutionary Changes Associated with the Progression of Human Immunodeficiency Virus type 1 Infection." J Virol 73(12):10489.

Sharp, P. M. and B. H. Hahn (2011). "Origins of HIV and the AIDS Pandemic." Cold Spring Harbor Perspectives in Medicine 1(1).

Shen, C., M. Ding, et al. (2012). "Evaluation of Cervical Mucosa in Transmission Bottleneck During Acute HIV-1 Infection Using a Cervical Tissue-based Organ Culture." PloS One 7(3): e32539.

Shimodaira, H. and M. Hasegawa (1999). "Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference." Molecular Biology and Evolution 16:1114.

Sickinger, E., G. Jonas, et al. (2008). "Performance Evaluation of the New Fully Automated Human Immunodeficiency Virus Antigen-antibody Combination Assay Designed for Blood Screening." Transfusion 48(4):584.

Siepel, A. C., A. L. Halpern, et al. (1995). "A Computer Program Designed to Screen Rapidly for HIV type 1 Intersubtype Recombinant Sequences." AIDS Research and Human Retroviruses 11(11):1413.

Siliciano, J. D., J. Kajdas, et al. (2003). "Long-term Follow-up Studies Confirm the Stability of the Latent Reservoir for HIV-1 in Resting CD4+ T Cells." Nature Medicine 9(6):727.

Soriano, V., C.-F. Perno, et al. (2009). "When and How to Use Maraviroc in HIV-Infected Patients." AIDS 23(18):2377.

Spiegelman, D., S. Schneeweiss, et al. (1997). "Measurement Error Correction for Logistic Regression Models with an "Alloyed Gold Standard"." American Journal of Epidemiology 145(2):184.

Stadeli, K. M. and D. D. Richman (2013). "Rates of Emergence of HIV Drug Resistance in Resource-limited Settings: a Systematic Review." Antivir Ther 18(1):115.

Stamatakis, A. (2014). "RAxML Version 8: a Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies." Bioinformatics(2014) May 1;30(9):1312.

Staszewski, S., V. Miller, et al. (1996). "Virological and Immunological Analysis of a Triple Combination Pilot Study with Loviride, Lamivudine and Zidovudine in HIV-1-infected Patients." AIDS 10(5):F1.

Stürmer, M., W. Preiser, et al. (2004). "Phylogenetic Analysis of HIV-1 Transmission: Pol Gene Sequences are Insufficient to Clarify True Relationships Between Patient Isolates." AIDS 18(16):2109.

Suligoi, B., C. Galli, et al. (2002). "Precision and Accuracy of a Procedure for Detecting Recent Human Immunodeficiency Virus Infections by Calculating the Antibody Avidity Index by an Automated Immunoassay-based Method." Journal of Clinical Microbiology 40(11):4015.

Sundquist, W. I. and H.-G. Kräusslich (2012). "HIV-1 Assembly, Budding, and Maturation." Cold Spring Harbor Perspectives in Medicine 2(7).

Tang, J. W. and D. Pillay (2004). "Transmission of HIV-1 Drug Resistance." Journal of Clinical Virology 30(1):1.

Tatt, I. D., K. L. Barlow, et al. (2004). "Surveillance of HIV-1 Subtypes Among Heterosexuals in England and Wales, 1997-2000." Journal of Acquired Immune Deficiency Syndromes 36(5):1092.

Terpe, K. (2013). "Overview of Thermostable DNA Polymerases for Classical PCR Applications: from Molecular and Biochemical Fundamentals to Commercial Systems." Applied Microbiology and Biotechnology 97(24):10243.

The UK Collaborative Group for HIV and STI Surveillance (2007). "Testing Times. HIV and Other Sexually Transmitted Infections in the United Kingdom: 2007."

U.K. Collaborative Group on HIV Drug Resistance (2012). "Time Trends in Drug Resistant HIV-1 Infections in the United Kingdom up to 2009: Multicentre Observational Study." BMJ 345:e5253.

UNAIDS (2013). "2013 Global Fact Sheet."

UNAIDS (2013). "Global Report: UNAIDS Report on the Global AIDS Epidemic 2013."

Usami, Y., S. Popov, et al. (2009). "The ESCRT Pathway and HIV-1 Budding." Biochemical Society Transactions 37(1):181.

Vallari, A., V. Holzmayer, et al. (2011). "Confirmation of Putative HIV-1 Group P in Cameroon." J Virol 85(3):1403.

Van't Wout, A. B., N. A. Kootstra, et al. (1994). "Macrophage-tropic Variants Initiate Human Immunodeficiency Virus type 1 Infection After Sexual, Parenteral, and Vertical Transmission." Journal of Clinical Investigation 94(5):2060.

Van Heuverswyn, F. and M. Peeters (2007). "The Origins of HIV and Implications for the Global Epidemic." Curr Infect Dis Rep 9(4):338.

Vandamme, A.-M. and O. G. Pybus (2013). "Viral Phylogeny in Court: the Unusual Case of the Valencian Anesthetist." BMC Biology 11(1):83.

Vandekerckhove, L. P. R., A. M. J. Wensing, et al. (2011). "European Guidelines on the Clinical Management of HIV-1 Tropism Testing." The Lancet Infectious Diseases 11(5):394.

Vilaseca, J., J. M. Arnau, et al. (1982). "Kaposi's Sarcoma and Toxoplasma Gondii Brain Abscess in a Spanish Homosexual." The Lancet 319(8271):572.

Volz, E. M., J. S. Koopman, et al. (2012). "Simple Epidemiological Dynamics Explain Phylogenetic Clustering of HIV from Patients with Recent Infection." PLoS Computational Biology 8(6):e1002552.

Wagner, G. A., M. E. Pacold, et al. (2013). "Incidence and Prevalence of Intrasubtype HIV-1 Dual Infection in At-Risk Men in the United States." Journal of Infectious Diseases: 2014 Apr 1;209(7):1032.

Wawer, M. J., R. H. Gray, et al. (2005). "Rates of HIV-1 Transmission per Coital Act, by Stage of HIV-1 Infection, in Rakai, Uganda." Journal of Infectious Diseases 191(9):1403.

Wawer, M. J., R. H. Gray, et al. (1998). "A Randomized, Community Trial of Intensive Sexually Transmitted Disease Control for AIDS Prevention, Rakai, Uganda." AIDS 12(10):1211.

Wawer, M. J., N. K. Sewankambo, et al. (1999). "Control of Sexually Transmitted Diseases for AIDS Prevention in Uganda: a Randomised Community Trial." The Lancet 353(9152):525.

Whelan, S., P. Liò, et al. (2001). "Molecular Phylogenetics: State-of-the-Art Methods for Looking into the Past." TRENDS in Genetics 17(5):262.

Wilke, C. O. (2005). "Quasispecies Theory in the Context of Population Genetics." BMC Evolutionary Biology 5(1):44.

Wong, J. K., M. Hezareh, et al. (1997). "Recovery of Replication-competent HIV Despite Prolonged Suppression of Plasma Viremia." Science 278(5341):1291.

Yang, Z. (1994). "Estimating the Pattern of Nucleotide Substitution." Journal of Molecular Evolution 39(1):105.

Yang, Z. (1996). "Among-site Rate Variation and its Impact on Phylogenetic Analyses." Trends in Ecology & Evolution 11(9):367.

Yang, Z. and B. Rannala (1997). "Bayesian Phylogenetic Inference using DNA Sequences: a Markov Chain Monte Carlo Method." Molecular Biology and Evolution 14(7):717.

Yang, Z. and B. Rannala (2012). "Molecular Phylogenetics: Principles and Practice." Nature Reviews Genetics 13(5):303.

Yarchoan, R., H. Mitsuya, et al. (1989). "In Vivo Activity Against HIV and Favorable Toxicity Profile of 2', 3'-dideoxyinosine." Science 245(4916):412.

Yarchoan, R., K. Weinhold, et al. (1986). "Administration of 3'-azido-3'-deoxythymidine, an Inhibitor of HTLV-III/LAV Replication, to Patients with AIDS or AIDS-related Complex." The Lancet 327(8481):575.

Yerly, S., T. Junier, et al. (2009). "The Impact of Transmission Clusters on Primary Drug Resistance in Newly Diagnosed HIV-1 Infection." AIDS 23(11):1415.

Yerly, S., L. Kaiser, et al. (1999). "Transmission of Antiretroviral-Drug-Resistant HIV-1 Variants." The Lancet 354(9180):729.

Yerly, S., S. Vora, et al. (2001). "Acute HIV Infection: Impact on the Spread of HIV and Transmission of Drug Resistance." AIDS 15(17):2287.

Youden, W. J. (1950). "Index for Rating Diagnostic Tests." Cancer 3(1):32.

Zagordi, O., A. Bhattacharya, et al. (2011). "ShoRAH: Estimating the Genetic Diversity of a Mixed Sample from Next-Generation Sequencing Data." BMC Bioinformatics 12(1):119.

Zagordi, O., M. Däumer, et al. (2012). "Read Length Versus Depth of Coverage for Viral Quasispecies Reconstruction." PloS One 7(10):e47046.

Zhang, L. Q., P. MacKenzie, et al. (1993). "Selection for Specific Sequences in the External Envelope Protein of Human Immunodeficiency Virus type 1 Upon Primary Infection." J Virol 67(6):3345.

Zhu, T., H. Mo, et al. (1993). "Genotypic and Phenotypic Characterization of HIV-1 Patients with Primary Infection." Science 261(5125):1179.

Appendices

Appendix 1: Table of primers for 454 deep-sequencing of HIV-1 *pol* and *env*

***Pol* amplicon**		
Primer A (LibA)	**MID**	**Forward *pol* gene template specific option 1 (NG_PinF2)**
CGTATCGCCTCCCTCGCGCCATCAG	ACGAGTGCGT	AGCATGACAAAAATCTTAGA
CGTATCGCCTCCCTCGCGCCATCAG	ACGCTCGACA	AGCATGACAAAAATCTTAGA
CGTATCGCCTCCCTCGCGCCATCAG	AGACGCACTC	AGCATGACAAAAATCTTAGA
CGTATCGCCTCCCTCGCGCCATCAG	AGCACTGTAG	AGCATGACAAAAATCTTAGA
CGTATCGCCTCCCTCGCGCCATCAG	ATCAGACACG	AGCATGACAAAAATCTTAGA
CGTATCGCCTCCCTCGCGCCATCAG	ATATCGCGAG	AGCATGACAAAAATCTTAGA
CGTATCGCCTCCCTCGCGCCATCAG	CGTGTCTCTA	AGCATGACAAAAATCTTAGA
CGTATCGCCTCCCTCGCGCCATCAG	CTCGCGTGTC	AGCATGACAAAAATCTTAGA
CGTATCGCCTCCCTCGCGCCATCAG	TCTCTATGCG	AGCATGACAAAAATCTTAGA
CGTATCGCCTCCCTCGCGCCATCAG	TGATACGTCT	AGCATGACAAAAATCTTAGA
CGTATCGCCTCCCTCGCGCCATCAG	CATAGTAGTG	AGCATGACAAAAATCTTAGA
CGTATCGCCTCCCTCGCGCCATCAG	CGAGAGATAC	AGCATGACAAAAATCTTAGA
CGTATCGCCTCCCTCGCGCCATCAG	ATACGACGTA	AGCATGACAAAAATCTTAGA
CGTATCGCCTCCCTCGCGCCATCAG	TCACGTACTA	AGCATGACAAAAATCTTAGA
CGTATCGCCTCCCTCGCGCCATCAG	CGTCTAGTAC	AGCATGACAAAAATCTTAGA
CGTATCGCCTCCCTCGCGCCATCAG	TCTACGTAGC	AGCATGACAAAAATCTTAGA
Primer B (LibA)	**MID**	**Reverse *pol* gene template specific option 1 (NG_PinR2)**
CTATGCGCCTTGCCAGCCCGCTCAG	ACGAGTGCGT	TGCCARTTCTARYTCTGCTTC
CTATGCGCCTTGCCAGCCCGCTCAG	ACGCTCGACA	TGCCARTTCTARYTCTGCTTC
CTATGCGCCTTGCCAGCCCGCTCAG	AGACGCACTC	TGCCARTTCTARYTCTGCTTC
CTATGCGCCTTGCCAGCCCGCTCAG	AGCACTGTAG	TGCCARTTCTARYTCTGCTTC
CTATGCGCCTTGCCAGCCCGCTCAG	ATCAGACACG	TGCCARTTCTARYTCTGCTTC
CTATGCGCCTTGCCAGCCCGCTCAG	ATATCGCGAG	TGCCARTTCTARYTCTGCTTC
CTATGCGCCTTGCCAGCCCGCTCAG	CGTGTCTCTA	TGCCARTTCTARYTCTGCTTC
CTATGCGCCTTGCCAGCCCGCTCAG	CTCGCGTGTC	TGCCARTTCTARYTCTGCTTC
CTATGCGCCTTGCCAGCCCGCTCAG	TCTCTATGCG	TGCCARTTCTARYTCTGCTTC
CTATGCGCCTTGCCAGCCCGCTCAG	TGATACGTCT	TGCCARTTCTARYTCTGCTTC
CTATGCGCCTTGCCAGCCCGCTCAG	CATAGTAGTG	TGCCARTTCTARYTCTGCTTC
CTATGCGCCTTGCCAGCCCGCTCAG	CGAGAGATAC	TGCCARTTCTARYTCTGCTTC
CTATGCGCCTTGCCAGCCCGCTCAG	ATACGACGTA	TGCCARTTCTARYTCTGCTTC
CTATGCGCCTTGCCAGCCCGCTCAG	TCACGTACTA	TGCCARTTCTARYTCTGCTTC

CTATGCGCCTTGCCAGCCCGCTCAG	CGTCTAGTAC	TGCCARTTCTARYTCTGCTTC
CTATGCGCCTTGCCAGCCCGCTCAG	TCTACGTAGC	TGCCARTTCTARYTCTGCTTC
Primer A (LibA)	**MID**	**Forward *pol* gene template specific option 2 (JA323)**
CGTATCGCCTCCCTCGCGCCATCAG	ACGAGTGCGT	TGGAAAGGATCACCAGCRATA
CGTATCGCCTCCCTCGCGCCATCAG	ACGCTCGACA	TGGAAAGGATCACCAGCRATA
CGTATCGCCTCCCTCGCGCCATCAG	AGACGCACTC	TGGAAAGGATCACCAGCRATA
CGTATCGCCTCCCTCGCGCCATCAG	AGCACTGTAG	TGGAAAGGATCACCAGCRATA
Primer B (LibA)	**MID**	**Reverse *pol* gene template specific option 2 (JA332)**
CTATGCGCCTTGCCAGCCCGCTCAG	ACGAGTGCGT	GCTGTACTGTCCATTTRTCAGGATG
CTATGCGCCTTGCCAGCCCGCTCAG	ACGCTCGACA	GCTGTACTGTCCATTTRTCAGGATG
CTATGCGCCTTGCCAGCCCGCTCAG	AGACGCACTC	GCTGTACTGTCCATTTRTCAGGATG
CTATGCGCCTTGCCAGCCCGCTCAG	AGCACTGTAG	GCTGTACTGTCCATTTRTCAGGATG
<u>Env amplicon</u>		
Primer A (LibA)	**MID**	**Forward env gene template specific option 1 (6955FQ)**
CGTATCGCCTCCCTCGCGCCATCAG	ACGAGTGCGT	CAGTACAATGYACACATGG
CGTATCGCCTCCCTCGCGCCATCAG	ACGCTCGACA	CAGTACAATGYACACATGG
CGTATCGCCTCCCTCGCGCCATCAG	AGACGCACTC	CAGTACAATGYACACATGG
CGTATCGCCTCCCTCGCGCCATCAG	AGCACTGTAG	CAGTACAATGYACACATGG
CGTATCGCCTCCCTCGCGCCATCAG	ATCAGACACG	CAGTACAATGYACACATGG
CGTATCGCCTCCCTCGCGCCATCAG	ATATCGCGAG	CAGTACAATGYACACATGG
CGTATCGCCTCCCTCGCGCCATCAG	CGTGTCTCTA	CAGTACAATGYACACATGG
CGTATCGCCTCCCTCGCGCCATCAG	CTCGCGTGTC	CAGTACAATGYACACATGG
CGTATCGCCTCCCTCGCGCCATCAG	TCTCTATGCG	CAGTACAATGYACACATGG
CGTATCGCCTCCCTCGCGCCATCAG	TGATACGTCT	CAGTACAATGYACACATGG
CGTATCGCCTCCCTCGCGCCATCAG	CATAGTAGTG	CAGTACAATGYACACATGG
CGTATCGCCTCCCTCGCGCCATCAG	CGAGAGATAC	CAGTACAATGYACACATGG
CGTATCGCCTCCCTCGCGCCATCAG	ATACGACGTA	CAGTACAATGYACACATGG
CGTATCGCCTCCCTCGCGCCATCAG	TCACGTACTA	CAGTACAATGYACACATGG
CGTATCGCCTCCCTCGCGCCATCAG	CGTCTAGTAC	CAGTACAATGYACACATGG
CGTATCGCCTCCCTCGCGCCATCAG	TCTACGTAGC	CAGTACAATGYACACATGG
Primer B (LibA)	**MID**	**Reverse env gene template specific option 1 (NG_EinR2)**

CTATGCGCCTTGCCAGCCCGCTCAG	ACGAGTGCGT	TAGAAAAATTCYCCTCYACAATTAAAR
CTATGCGCCTTGCCAGCCCGCTCAG	ACGCTCGACA	TAGAAAAATTCYCCTCYACAATTAAAR
CTATGCGCCTTGCCAGCCCGCTCAG	AGACGCACTC	TAGAAAAATTCYCCTCYACAATTAAAR
CTATGCGCCTTGCCAGCCCGCTCAG	AGCACTGTAG	TAGAAAAATTCYCCTCYACAATTAAAR
CTATGCGCCTTGCCAGCCCGCTCAG	ATCAGACACG	TAGAAAAATTCYCCTCYACAATTAAAR
CTATGCGCCTTGCCAGCCCGCTCAG	ATATCGCGAG	TAGAAAAATTCYCCTCYACAATTAAAR
CTATGCGCCTTGCCAGCCCGCTCAG	CGTGTCTCTA	TAGAAAAATTCYCCTCYACAATTAAAR
CTATGCGCCTTGCCAGCCCGCTCAG	CTCGCGTGTC	TAGAAAAATTCYCCTCYACAATTAAAR
CTATGCGCCTTGCCAGCCCGCTCAG	TCTCTATGCG	TAGAAAAATTCYCCTCYACAATTAAAR
CTATGCGCCTTGCCAGCCCGCTCAG	TGATACGTCT	TAGAAAAATTCYCCTCYACAATTAAAR
CTATGCGCCTTGCCAGCCCGCTCAG	CATAGTAGTG	TAGAAAAATTCYCCTCYACAATTAAAR
CTATGCGCCTTGCCAGCCCGCTCAG	CGAGAGATAC	TAGAAAAATTCYCCTCYACAATTAAAR
CTATGCGCCTTGCCAGCCCGCTCAG	ATACGACGTA	TAGAAAAATTCYCCTCYACAATTAAAR
CTATGCGCCTTGCCAGCCCGCTCAG	TCACGTACTA	TAGAAAAATTCYCCTCYACAATTAAAR
CTATGCGCCTTGCCAGCCCGCTCAG	CGTCTAGTAC	TAGAAAAATTCYCCTCYACAATTAAAR
CTATGCGCCTTGCCAGCCCGCTCAG	TCTACGTAGC	TAGAAAAATTCYCCTCYACAATTAAAR
Primer A (LibA)	**MID**	**Forward env gene template specific option 2 (Roz_sense)**
CGTATCGCCTCCCTCGCGCCATCAG	ACGAGTGCGT	TGGCAGTCTAGCAGAAGAAG
CGTATCGCCTCCCTCGCGCCATCAG	ACGCTCGACA	TGGCAGTCTAGCAGAAGAAG
CGTATCGCCTCCCTCGCGCCATCAG	AGACGCACTC	TGGCAGTCTAGCAGAAGAAG
CGTATCGCCTCCCTCGCGCCATCAG	AGCACTGTAG	TGGCAGTCTAGCAGAAGAAG
Primer B (LibA)	**MID**	**Reverse env gene template specific option 2 (Roz_anti)**
CTATGCGCCTTGCCAGCCCGCTCAG	ACGAGTGCGT	CTGGGTCCCCTCCTGAGG
CTATGCGCCTTGCCAGCCCGCTCAG	ACGCTCGACA	CTGGGTCCCCTCCTGAGG
CTATGCGCCTTGCCAGCCCGCTCAG	AGACGCACTC	CTGGGTCCCCTCCTGAGG
CTATGCGCCTTGCCAGCCCGCTCAG	AGCACTGTAG	CTGGGTCCCCTCCTGAGG

Appendix 2: Perl scripts (available in electronic format on request)

findclusterdistances.pl

```
#!usr/bin/perl -w

# Program for finding clusters with a support value greater than a specified amount and for finding the
distances between all nodes in such clusters
use Math::BigInt;
use Math::BigFloat;
use Bio::TreeIO;
use Scalar::Util qw(looks_like_number);

my $input_file;
my $output_file;

if (defined($ARGV[0])) {
  chomp($tree_input_file = $ARGV[0]);
} else {
  print "Please enter a newick formatted tree file (with full path if not in this directory): ";
  chomp($tree_input_file = <>);
}

if (defined($ARGV[1])) {
  chomp($sequences_input_file = $ARGV[1]);
} else {
  print "Please enter a fasta formatted sequence file (with full path if not in this directory): ";
  chomp($sequences_input_file = <>);
}

my $support_score = 0;
my $lower_cluster_size_filter = 0;
my $upper_cluster_size_filter = 2000;

system ("rm -rf
$sequences_input_file.$support_score.$lower_cluster_size_filter.$upper_cluster_size_filter.findclusteroutpu
t.txt");
system ("rm -rf $sequences_input_file.temp_R.txt");
system ("rm -rf $sequences_input_file.rscript.R");
system ("rm -rf $sequences_input_file.R_distances.txt");
system ("rm -rf $sequences_input_file.rscript.Rout");

print "\n\nLoading tree...";

my $input = Bio::TreeIO->new(-file => $tree_input_file, -format=>'newick');

################## CREATING LIST OF SEQUENCES INVOLVED IN CLUSTERS SECTION ##################

my @nodearray;
my $nodecount = 0;

while($tree = $input->next_tree){
         print "\n\nProcessing nodes...";
         for my $node ($tree->get_nodes){
                  if(!$node->is_Leaf){
                           if(looks_like_number($node->id)){
                                    if($node->id >= ($support_score/100)){
                                             my @predistancearray;
                                             @descendents = $node->get_all_Descendents;
                                             my $descendentcount = 0;
                                             foreach $descendent (@descendents){
                                                      if($descendent->is_Leaf){
                                                               # @predistancearray is an array to store all
leaves in a particular cluster for future calculation of their distances from each other
                                                               $predistancearray[$descendentcount] =
$descendent->id;
                                                               $descendentcount++;
                                                      }
                                             }
                                             # This if conditional is to filter out massive clusters
                                             if (@predistancearray > $lower_cluster_size_filter &&
@predistancearray <= $upper_cluster_size_filter){
                                                      # $nodearray is an array of references to arrays with
all leaves from a particular cluster
                                                      $nodearray[$nodecount] = [@predistancearray];
                                                      $nodecount++;
                                             }
                                    }
                           }
                  }
         }
}

# Loop to remove duplicates from cluster list because some sequences will be in clusters within clusters
my %unique_cluster_ids = ();
```

```
for $arrayref (@nodearray){
        for $id (@$arrayref){
                $unique_cluster_ids{$id} = 1;
        }
}

################# SELECT SEQUENCES SECTION #################

open(SEQUENCES, $sequences_input_file);

my @identifiers = keys %unique_cluster_ids;

my %sequences;
my $switch = "start";
my $identifier = "";
my $goodformat = "true";
my $index = 1;

while(my $fileline = <SEQUENCES>){
        if($fileline =~ /^>(\S*)/ && $switch eq "start"){
                $identifier = $1;
                $switch = "identifier";
        }elsif($fileline =~ /^>(\S*)/ && $switch eq "sequence"){
                $identifier = $1;
                $switch = "identifier";
        }elsif($fileline !~ /^>\S*/ && $switch eq "identifier"){
                $sequences{$identifier} = $fileline;
                $switch = "sequence";
        }else{
                $goodformat = "false";
                if($switch eq "start"){
                        print "\nSequence file not in correct fasta format around line $index.\n";
                }else{
                        print "\nSequence file not in correct fasta format around sequence
$identifier.\n";
                }
                die();
        }
        $index++;
}

if($goodformat){
        print "\n\nInput sequence file in good fasta format...\n\n";
}

for $i (0..$#nodearray){
        my $currentclustercount = $i+1;
        print "processing cluster $currentclustercount of $nodecount\n";
        # get reference to the $ith array of leaf nodes, i.e. a particular cluster
        $arrayref = $nodearray[$i];
        # get length of array, which is how many sequences are in the particular cluster - !!! REMEMBER
that arrays start at 0, e.g. $n == 2 will pull out all clusters of size 3 NOT size 2 (0,1,2) !!!
        $n = @$arrayref - 1;
        open(OUTPUT, ">$sequences_input_file.temp_R.txt");
        # $j will be the element in the $ith array up to the length of $i array
        for $j (0..$n){
                #foreach my $array (@identifiers){
                if($sequences{$nodearray[$i][$j]}){
                        print OUTPUT ">$nodearray[$i][$j]\n$sequences{$nodearray[$i][$j]}";
                }else{
                        print "Sequence $nodearray[$i][$j] is not present in sequence file.\n";
                }
        }

        close OUTPUT;

################# R DISTANCE CALCULATION SECTION #################

        open(RSCRIPT, ">$sequences_input_file.rscript.R");
        print RSCRIPT "library(ape)"."\n";
        print RSCRIPT "seqs<-read.dna('$sequences_input_file.temp_R.txt',format='fasta')"."\n";
        print RSCRIPT "dist<-dist.dna(seqs,model='raw',pairwise.deletion=T,as.matrix=T)"."\n";
        print RSCRIPT "write.table(dist,file='$sequences_input_file.R_distances.txt',quote=F)"."\n";
        close RSCRIPT;

        system ("R CMD BATCH $sequences_input_file.rscript.R");


################# R DISTANCE MATRIX INPUT SECTION #################

        open(RSCRIPTINPUT, "$sequences_input_file.R_distances.txt");

        my @AoA = ();

        # Loop to split each line of the distance matrix by " " and store in an array, then push this into
an array of arrays
        while(my $fileline = <RSCRIPTINPUT>){
                my @tmp = split(" ",$fileline);
```

```
                push(@AoA, [@tmp]);
        }

        my $count = 0;
        my %hash_of_arrays = ();
        my %hash_of_column_positions = ();

        # This foreach loop stores the distances for one particular sequence to all other sequences in a
hash array
        # with the key the sequence name (which should always be the first position in the matrix output
by R
        # The hash_of_columns is to store the column that a particular sequence's distance will be from
the sequence
        # row being dealt with, i.e. the first sequence row is for x, and is a row of distances
corresponding to the
        # other sequences,y,z... the second row is for sequence y, and its first distance position will be
for the
        # distance to sequence x
        foreach my $array(@AoA){

                ###############################################################################
                # This line uses the first element in each array as the name of the sequence, only #
                # to be employed when using write.table in R to write out a matrix. This will      #
                # overwrite the first line of the matrix, which will just be a line of sequence    #
                # identifiers - but $count will have already been incremented by one because of    #
                # the first line of the R output                                                   #
                ###############################################################################
                $hash_of_arrays{@$array[0]} = [@$array];
                $hash_of_column_positions{@$array[0]} = $count;
                $count++;
        }


################## AVERAGE PAIRWISE DISTANCE CALCULATION SECTION #################

        open(FINDCLUSTEROUTPUT,
">>$sequences_input_file.$support_score.$lower_cluster_size_filter.$upper_cluster_size_filter.findclusterou
tput.txt");

        # The following variable is created for each cluster, and is a non-duplicate list of all nodes in
the cluster - this is sorted below and used as a unique identifier for that cluster
        my %unique_cluster_ids_hash = ();
        my $sum_distances = 0;
        # This variable is to account for some distances that are unmeasurable ("NA") in the matrix and
cannot be summed (because they do not have overlapping sequence, i.e. one has gaps aligned to the other's
sequence and vice versa)
        my $unmeasurable_distances_count = 0;
        for $j (0..$n){
                $m = $n - $j;
                for $k (0..$m){
                        $l = $k + $j;
                        my $node1 = $nodearray[$i][$j];
                        my $node2 = $nodearray[$i][$l];
                        my $distance = $hash_of_arrays{$node1}[$hash_of_column_positions{$node2}];
                        print FINDCLUSTEROUTPUT "Distance between node $node1 and node $node2 is
$distance\n";
                        if(looks_like_number($distance)){
                                $sum_distances += $distance;
                        }else{
                                # Add a count to $unmeasurable_distances_count to reduce nCr by one
because one pairwise distance was unmeasurable ("NA") - this happens when two sequences do not overlap at
all in alignedsequences.fas
                                $unmeasurable_distances_count++;
                        }
                        $unique_cluster_ids_hash{$node1} = 1;
                        $unique_cluster_ids_hash{$node2} = 1;
                }
        }
        # Use bcom subroutine to use factorials to calculate nCr
        my $nCr = bcomb(($n+1), 2);
        # Reduce nCr by the number of pairwise distances that were unmeasurable above
        $nCr = $nCr-$unmeasurable_distances_count;
        # Use divide subroutine to divide with BigFloat
        $mean_distances = divide($sum_distances, $nCr);
        print FINDCLUSTEROUTPUT "mean of cluster = $mean_distances:";
        foreach $unique_cluster_id (sort(keys %unique_cluster_ids_hash)){
                print FINDCLUSTEROUTPUT " "."$unique_cluster_id";
        }
        print FINDCLUSTEROUTPUT "\n\n";
        close FINDCLUSTEROUTPUT;

        system ("rm -rf $sequences_input_file.temp_R.txt");
        system ("rm -rf $sequences_input_file.rscript.R");
        system ("rm -rf $sequences_input_file.R_distances.txt");
        system ("rm -rf $sequences_input_file.rscript.Rout");
}
```

```
print "\n\nSee
$sequences_input_file.$support_score.$lower_cluster_size_filter.$upper_cluster_size_filter.findclusteroutpu
t.txt for cluster average pairwise genetic distances.\n\n";

sub divide{
        my ($nom, $denom) = @_;
        $nom = Math::BigFloat->new($nom);
        return $nom->copy()->bdiv($denom);
}

sub bcomb{
        my ($n, $r) = @_;
        $n = Math::BigInt->new($n);
        $r = Math::BigInt->new($r);
        my $k = $n - $r;
        return $n->bfac()/($r->bfac()*$k->bfac());
}
```

recentestablishedsorter_adapted_for_infector_assignment.pl

```
#!usr/bin/perl -w

use strict;
use warnings;
use List::Util qw(shuffle);

#
# File to find number of recent and established infections within a cluster that could have seeded other
cluster infections
#

if (defined($ARGV[0])) {
  open(INPUT1, $ARGV[0]);
} else {
  print 'No recent/established status file of the form "sequenceId sampleDate recent/establishedStatus"\n';
        my $input = <>;
  open(INPUT1, $input);
}

if (defined($ARGV[1])) {
  open(INPUT2, $ARGV[1]);
} else {
  print "No treatment experience file, please enter one:\n";
        my $input = <>;
  open(INPUT2, $input);
}

if (defined($ARGV[2])) {
  open(INPUT3, $ARGV[2]);
} else {
  print "No clusters file, please enter one:\n";
        my $input = <>;
  open(INPUT3, $input);
}

my $min_analysis_size;
if (defined($ARGV[3])) {
  if($ARGV[3] =~ /\d+/){
    $min_analysis_size = $ARGV[3]-1;
  } else {
    die ("No minimum cluster size specified\n");
  }
} else {
   die ("No minimum cluster size specified\n");
}

my $recent_phase_length;
if (defined($ARGV[4])) {
  if($ARGV[4] =~ /\d+/){
    $recent_phase_length = $ARGV[4];
  } else {
    die ("No recent phase length specified as final argument\n");
  }
} else {
   die ("No recent phase length specified as final argument\n");
}

#my $output = "recentestablishedsorter_random.txt";
#open(OUTPUT, ">$output");

my %sequence_infection_status_hash = ();
my %sequence_sampledate_hash = ();
my %sequence_therapy_status_hash = ();
my %sequence_drug_resistance_status_hash = ();
my %hash_of_cluster_arrays = ();
my %hash_of_cluster_lengths = ();
my %hash_upper_dates = ();
my %hash_lower_dates = ();
my %duplicate_recents_infection_dates = ();
my %duplicate_recents_infection_dates_experienced = ();
my %duplicate_recents_infection_dates_notclassified = ();
my %duplicate_recents_infection_dates_naive = ();

my $infection_window_for_recents = $recent_phase_length;

while(my $fileline = <INPUT1>){
        chomp($fileline);
        # INPUT1 should be a tab delimited list of sequence id followed by dbsample date as a number
followed by recent or established status followed by drug resistance (D for drug resistance, S for drug
sensitive)
        my @fileline = split(/\t/,$fileline);
        # Use sequence id as key, and status as value
        if(!$sequence_infection_status_hash{$fileline[0]}){
                $sequence_sampledate_hash{$fileline[0]} = $fileline[1];
```

```
                $sequence_infection_status_hash{$fileline[0]} = $fileline[2];
                $sequence_drug_resistance_status_hash{$fileline[0]} = $fileline[3];
        }else{
                die "Sequence $fileline[0] used more than once!\n";
        }
}

while(my $fileline = <INPUT2>){
        chomp($fileline);
        # INPUT2 should be a tab delimited list of sequence id followed by therapy status
        my @fileline = split(/,/,$fileline);
        # Use sequence id as key, and status as value
        if(!$sequence_therapy_status_hash{$fileline[0]}){
                $sequence_therapy_status_hash{$fileline[0]} = $fileline[1];
        }else{
                die "Sequence $fileline[0] used more than once!\n";
        }
}

while(my $fileline = <INPUT3>){
        if($fileline =~ /^\s\d/){
                chomp($fileline);
                # USE // WITH split() NOT '' with this form of cluster id - using // means first element
is empty, because there's a space at the beginning of the line e.g. " 101253 112376 113698"
                my @clusterarray = split(/\s/, $fileline);
                splice(@clusterarray,0,1);
                $hash_of_cluster_arrays{$fileline} = [@clusterarray];
                $hash_of_cluster_lengths{$fileline} = @clusterarray;
        }
}

##Loop to filter out non-recents
my %hash_of_recents_cluster_arrays = ();
my $number_of_suitable_clusters = 0;
foreach my $unique_cluster_id (sort keys %hash_of_cluster_arrays){
        my @recents_array = ();

############################################################
# Added in a filter to select for treatment naive patients #
############################################################

        foreach my $sequence1 (@{$hash_of_cluster_arrays{$unique_cluster_id}}){
                if($sequence_infection_status_hash{$sequence1} eq "recent" &&
$sequence_therapy_status_hash{$sequence1} eq '"Naive"'){
                        push @recents_array, $sequence1;
                }
        }
        #Must be more than a specified number of recents in the cluster for the ordersorter section to
work
        if(scalar @recents_array > $min_analysis_size){
                $number_of_suitable_clusters++;
                $hash_of_recents_cluster_arrays{$unique_cluster_id} = [@recents_array];
        }
}

my $iterations = 10;
my $count = 0;
my @temp_array = ();
my @recents_in_cluster_array = ();
my @cluster_number_of_recents_array = ();
my @cluster_number_of_experienced_recents_array = ();
my @cluster_number_of_naive_recents_array = ();
my @cluster_number_of_notclassified_recents_array = ();
my @cluster_number_of_drug_resistance_recents_array = ();
my @cluster_number_of_experienced_drug_resistance_recents_array = ();
my @cluster_number_of_naive_drug_resistance_recents_array = ();
my @cluster_number_of_notclassified_drug_resistance_recents_array = ();
my @cluster_number_of_establisheds_array = ();
my @cluster_number_of_experienced_establisheds_array = ();
my @cluster_number_of_naive_establisheds_array = ();
my @cluster_number_of_notclassified_establisheds_array = ();
my @cluster_number_of_drug_resistance_establisheds_array = ();
my @cluster_number_of_experienced_drug_resistance_establisheds_array = ();
my @cluster_number_of_naive_drug_resistance_establisheds_array = ();
my @cluster_number_of_notclassified_drug_resistance_establisheds_array = ();
my $recents_in_cluster_array = 0;
my $cluster_number_of_recents_array = 0;
my $cluster_number_of_experienced_recents_array = 0;
my $cluster_number_of_naive_recents_array = 0;
my $cluster_number_of_notclassified_recents_array = 0;
my $cluster_number_of_drug_resistance_recents_array = 0;
my $cluster_number_of_experienced_drug_resistance_recents_array = 0;
my $cluster_number_of_naive_drug_resistance_recents_array = 0;
my $cluster_number_of_notclassified_drug_resistance_recents_array = 0;
my $cluster_number_of_establisheds_array = 0;
my $cluster_number_of_experienced_establisheds_array = 0;
my $cluster_number_of_naive_establisheds_array = 0;
my $cluster_number_of_notclassified_establisheds_array = 0;
```

```
my $cluster_number_of_drug_resistance_establisheds_array = 0;
my $cluster_number_of_experienced_drug_resistance_establisheds_array = 0;
my $cluster_number_of_naive_drug_resistance_establisheds_array = 0;
my $cluster_number_of_notclassified_drug_resistance_establisheds_array = 0;

if(keys %hash_of_recents_cluster_arrays > 0){
        my $random1 = 0;
        my $random2 = 0;
        my $ordered1 = 0;
        my $ordered2 = 0;
        my $fake1 = 0;
        my $fake2 = 0;
        my $fake_random1 = 0;
        my $fake_random2 = 0;
        ##Loop to randomly assign infectors to infectees
        while($count < $iterations){
                foreach my $unique_cluster_id (sort keys %hash_of_recents_cluster_arrays){
                        my %random_infector_hash = ();
                        #Date order sequences in cluster to see if there is more than one sequence with
the earliest date (i.e. the same date)
                        my %dbsample_dates_hash = ();
                        foreach my $sequence1 (@{$hash_of_recents_cluster_arrays{$unique_cluster_id}}){
                                $dbsample_dates_hash{$sequence1} = $sequence_sampledate_hash{$sequence1};
                        }
                        my @cluster_members_sorted_by_dbsample_date = ();
                        @cluster_members_sorted_by_dbsample_date = sort{$dbsample_dates_hash{$a} cmp
$dbsample_dates_hash{$b}} keys %dbsample_dates_hash;
                        %random_infector_hash =
random_infector_assignment(\@cluster_members_sorted_by_dbsample_date, \%dbsample_dates_hash,
$unique_cluster_id);
                        @temp_array =
@{ordersorter($unique_cluster_id,\%random_infector_hash,\%sequence_sampledate_hash)};
                        push @recents_in_cluster_array, $temp_array[0];
                        push @cluster_number_of_recents_array, $temp_array[1];
                        push @cluster_number_of_experienced_recents_array, $temp_array[2];
                        push @cluster_number_of_naive_recents_array, $temp_array[3];
                        push @cluster_number_of_notclassified_recents_array, $temp_array[4];
                        push @cluster_number_of_drug_resistance_recents_array, $temp_array[5];
                        push @cluster_number_of_experienced_drug_resistance_recents_array,
$temp_array[6];
                        push @cluster_number_of_naive_drug_resistance_recents_array, $temp_array[7];
                        push @cluster_number_of_notclassified_drug_resistance_recents_array,
$temp_array[8];
                        push @cluster_number_of_establisheds_array, $temp_array[9];
                        push @cluster_number_of_experienced_establisheds_array, $temp_array[10];
                        push @cluster_number_of_naive_establisheds_array, $temp_array[11];
                        push @cluster_number_of_notclassified_establisheds_array, $temp_array[12];
                        push @cluster_number_of_drug_resistance_establisheds_array, $temp_array[13];
                        push @cluster_number_of_experienced_drug_resistance_establisheds_array,
$temp_array[14];
                        push @cluster_number_of_naive_drug_resistance_establisheds_array,
$temp_array[15];
                        push @cluster_number_of_notclassified_drug_resistance_establisheds_array,
$temp_array[16];
                        $recents_in_cluster_array = $recents_in_cluster_array + $temp_array[0];
                        $cluster_number_of_recents_array = $cluster_number_of_recents_array +
$temp_array[1];
                        $cluster_number_of_experienced_recents_array =
$cluster_number_of_experienced_recents_array + $temp_array[2];
                        $cluster_number_of_naive_recents_array = $cluster_number_of_naive_recents_array +
$temp_array[3];
                        $cluster_number_of_notclassified_recents_array =
$cluster_number_of_notclassified_recents_array + $temp_array[4];
                        $cluster_number_of_drug_resistance_recents_array =
$cluster_number_of_drug_resistance_recents_array + $temp_array[5];
                        $cluster_number_of_experienced_drug_resistance_recents_array =
$cluster_number_of_experienced_drug_resistance_recents_array + $temp_array[6];
                        $cluster_number_of_naive_drug_resistance_recents_array =
$cluster_number_of_naive_drug_resistance_recents_array + $temp_array[7];
                        $cluster_number_of_notclassified_drug_resistance_recents_array =
$cluster_number_of_notclassified_drug_resistance_recents_array + $temp_array[8];
                        $cluster_number_of_establisheds_array = $cluster_number_of_establisheds_array +
$temp_array[9];
                        $cluster_number_of_experienced_establisheds_array =
$cluster_number_of_experienced_establisheds_array + $temp_array[10];
                        $cluster_number_of_naive_establisheds_array =
$cluster_number_of_naive_establisheds_array + $temp_array[11];
                        $cluster_number_of_notclassified_establisheds_array =
$cluster_number_of_notclassified_establisheds_array + $temp_array[12];
                        $cluster_number_of_drug_resistance_establisheds_array =
$cluster_number_of_drug_resistance_establisheds_array + $temp_array[13];
                        $cluster_number_of_experienced_drug_resistance_establisheds_array =
$cluster_number_of_experienced_drug_resistance_establisheds_array + $temp_array[14];
                        $cluster_number_of_naive_drug_resistance_establisheds_array =
$cluster_number_of_naive_drug_resistance_establisheds_array + $temp_array[15];
                        $cluster_number_of_notclassified_drug_resistance_establisheds_array =
$cluster_number_of_notclassified_drug_resistance_establisheds_array + $temp_array[16];
                }
```

```
                $count++;
        }

        $random1 = $recents_in_cluster_array/$iterations;
        print "random:total_of_recent_infection_events_in_clusters:$random1;";
        $random2 = $cluster_number_of_recents_array/$iterations;
        print "random:total_of_recent_phase_infections:$random2;";

        ##Loop to date order assign infectors to infectees
        @temp_array = ();
        @recents_in_cluster_array = ();
        @cluster_number_of_recents_array = ();
        @cluster_number_of_experienced_recents_array = ();
        @cluster_number_of_naive_recents_array = ();
        @cluster_number_of_notclassified_recents_array = ();
        @cluster_number_of_drug_resistance_recents_array = ();
        @cluster_number_of_experienced_drug_resistance_recents_array = ();
        @cluster_number_of_naive_drug_resistance_recents_array = ();
        @cluster_number_of_notclassified_drug_resistance_recents_array = ();
        @cluster_number_of_establisheds_array = ();
        @cluster_number_of_experienced_establisheds_array = ();
        @cluster_number_of_naive_establisheds_array = ();
        @cluster_number_of_notclassified_establisheds_array = ();
        @cluster_number_of_drug_resistance_establisheds_array = ();
        @cluster_number_of_experienced_drug_resistance_establisheds_array = ();
        @cluster_number_of_naive_drug_resistance_establisheds_array = ();
        @cluster_number_of_notclassified_drug_resistance_establisheds_array = ();
        $recents_in_cluster_array = 0;
        $cluster_number_of_recents_array = 0;
        $cluster_number_of_experienced_recents_array = 0;
        $cluster_number_of_naive_recents_array = 0;
        $cluster_number_of_notclassified_recents_array = 0;
        $cluster_number_of_drug_resistance_recents_array = 0;
        $cluster_number_of_experienced_drug_resistance_recents_array = 0;
        $cluster_number_of_naive_drug_resistance_recents_array = 0;
        $cluster_number_of_notclassified_drug_resistance_recents_array = 0;
        $cluster_number_of_establisheds_array = 0;
        $cluster_number_of_experienced_establisheds_array = 0;
        $cluster_number_of_naive_establisheds_array = 0;
        $cluster_number_of_notclassified_establisheds_array = 0;
        $cluster_number_of_drug_resistance_establisheds_array = 0;
        $cluster_number_of_experienced_drug_resistance_establisheds_array = 0;
        $cluster_number_of_naive_drug_resistance_establisheds_array = 0;
        $cluster_number_of_notclassified_drug_resistance_establisheds_array = 0;
        ##Ordered assignment section
        foreach my $unique_cluster_id (sort(keys %hash_of_recents_cluster_arrays)){
                my %ordered_infector_hash = ();
                my %dbsample_dates_hash = ();
                foreach my $sequence1 (@{$hash_of_recents_cluster_arrays{$unique_cluster_id}}){
                        $dbsample_dates_hash{$sequence1} = $sequence_sampledate_hash{$sequence1};
                }
                my @cluster_members_sorted_by_dbsample_date = ();
                @cluster_members_sorted_by_dbsample_date = sort{$dbsample_dates_hash{$a} cmp
$dbsample_dates_hash{$b}} keys %dbsample_dates_hash;
                my $array_count = 0;
                foreach (@cluster_members_sorted_by_dbsample_date){
                        if($array_count < scalar @cluster_members_sorted_by_dbsample_date && $array_count
> 0){

        $ordered_infector_hash{$cluster_members_sorted_by_dbsample_date[$array_count]} =
$cluster_members_sorted_by_dbsample_date[$array_count - 1];
                        }
                        $array_count++;
                }
                @temp_array =
@{ordersorter($unique_cluster_id,\%ordered_infector_hash,\%sequence_sampledate_hash)};
                push @recents_in_cluster_array, $temp_array[0];
                push @cluster_number_of_recents_array, $temp_array[1];
                push @cluster_number_of_experienced_recents_array, $temp_array[2];
                push @cluster_number_of_naive_recents_array, $temp_array[3];
                push @cluster_number_of_notclassified_recents_array, $temp_array[4];
                push @cluster_number_of_drug_resistance_recents_array, $temp_array[5];
                push @cluster_number_of_experienced_drug_resistance_recents_array, $temp_array[6];
                push @cluster_number_of_naive_drug_resistance_recents_array, $temp_array[7];
                push @cluster_number_of_notclassified_drug_resistance_recents_array, $temp_array[8];
                push @cluster_number_of_establisheds_array, $temp_array[9];
                push @cluster_number_of_experienced_establisheds_array, $temp_array[10];
                push @cluster_number_of_naive_establisheds_array, $temp_array[11];
                push @cluster_number_of_notclassified_establisheds_array, $temp_array[12];
                push @cluster_number_of_drug_resistance_establisheds_array, $temp_array[13];
                push @cluster_number_of_experienced_drug_resistance_establisheds_array, $temp_array[14];
                push @cluster_number_of_naive_drug_resistance_establisheds_array, $temp_array[15];
                push @cluster_number_of_notclassified_drug_resistance_establisheds_array, $temp_array[16];
                $recents_in_cluster_array = $recents_in_cluster_array + $temp_array[0];
                $cluster_number_of_recents_array = $cluster_number_of_recents_array + $temp_array[1];
                $cluster_number_of_experienced_recents_array =
$cluster_number_of_experienced_recents_array + $temp_array[2];
```

```
                $cluster_number_of_naive_recents_array = $cluster_number_of_naive_recents_array +
$temp_array[3];
                $cluster_number_of_notclassified_recents_array =
$cluster_number_of_notclassified_recents_array + $temp_array[4];
                $cluster_number_of_drug_resistance_recents_array =
$cluster_number_of_drug_resistance_recents_array + $temp_array[5];
                $cluster_number_of_experienced_drug_resistance_recents_array =
$cluster_number_of_experienced_drug_resistance_recents_array + $temp_array[6];
                $cluster_number_of_naive_drug_resistance_recents_array =
$cluster_number_of_naive_drug_resistance_recents_array + $temp_array[7];
                $cluster_number_of_notclassified_drug_resistance_recents_array =
$cluster_number_of_notclassified_drug_resistance_recents_array + $temp_array[8];
                $cluster_number_of_establisheds_array = $cluster_number_of_establisheds_array +
$temp_array[9];
                $cluster_number_of_experienced_establisheds_array =
$cluster_number_of_experienced_establisheds_array + $temp_array[10];
                $cluster_number_of_naive_establisheds_array = $cluster_number_of_naive_establisheds_array
+ $temp_array[11];
                $cluster_number_of_notclassified_establisheds_array =
$cluster_number_of_notclassified_establisheds_array + $temp_array[12];
                $cluster_number_of_drug_resistance_establisheds_array =
$cluster_number_of_drug_resistance_establisheds_array + $temp_array[13];
                $cluster_number_of_experienced_drug_resistance_establisheds_array =
$cluster_number_of_experienced_drug_resistance_establisheds_array + $temp_array[14];
                $cluster_number_of_naive_drug_resistance_establisheds_array =
$cluster_number_of_naive_drug_resistance_establisheds_array + $temp_array[15];
                $cluster_number_of_notclassified_drug_resistance_establisheds_array =
$cluster_number_of_notclassified_drug_resistance_establisheds_array + $temp_array[16];
        }

        $ordered1 = $recents_in_cluster_array;
        print "ordered:total_of_recent_infection_events_in_clusters:$ordered1;";
        $ordered2 = $cluster_number_of_recents_array;
        print "ordered:total_of_recent_phase_infections:$ordered2;";

        my @fake_dbsample_dates_temp_array = ();
        my @fake_dbsample_dates_recents_in_cluster_array = ();
        my @fake_dbsample_dates_cluster_number_of_recents_array = ();
        my @fake_dbsample_dates_cluster_number_of_experienced_recents_array = ();
        my @fake_dbsample_dates_cluster_number_of_naive_recents_array = ();
        my @fake_dbsample_dates_cluster_number_of_notclassified_recents_array = ();
        my @fake_dbsample_dates_cluster_number_of_drug_resistance_recents_array = ();
        my @fake_dbsample_dates_cluster_number_of_experienced_drug_resistance_recents_array = ();
        my @fake_dbsample_dates_cluster_number_of_naive_drug_resistance_recents_array = ();
        my @fake_dbsample_dates_cluster_number_of_notclassified_drug_resistance_recents_array = ();
        my @fake_dbsample_dates_cluster_number_of_establisheds_array = ();
        my @fake_dbsample_dates_cluster_number_of_experienced_establisheds_array = ();
        my @fake_dbsample_dates_cluster_number_of_naive_establisheds_array = ();
        my @fake_dbsample_dates_cluster_number_of_notclassified_establisheds_array = ();
        my @fake_dbsample_dates_cluster_number_of_drug_resistance_establisheds_array = ();
        my @fake_dbsample_dates_cluster_number_of_experienced_drug_resistance_establisheds_array = ();
        my @fake_dbsample_dates_cluster_number_of_naive_drug_resistance_establisheds_array = ();
        my @fake_dbsample_dates_cluster_number_of_notclassified_drug_resistance_establisheds_array = ();
        my $fake_dbsample_dates_temp_array = 0;
        my $fake_dbsample_dates_recents_in_cluster_array = 0;
        my $fake_dbsample_dates_cluster_number_of_recents_array = 0;
        my $fake_dbsample_dates_cluster_number_of_experienced_recents_array = 0;
        my $fake_dbsample_dates_cluster_number_of_naive_recents_array = 0;
        my $fake_dbsample_dates_cluster_number_of_notclassified_recents_array = 0;
        my $fake_dbsample_dates_cluster_number_of_drug_resistance_recents_array = 0;
        my $fake_dbsample_dates_cluster_number_of_experienced_drug_resistance_recents_array = 0;
        my $fake_dbsample_dates_cluster_number_of_naive_drug_resistance_recents_array = 0;
        my $fake_dbsample_dates_cluster_number_of_notclassified_drug_resistance_recents_array = 0;
        my $fake_dbsample_dates_cluster_number_of_establisheds_array = 0;
        my $fake_dbsample_dates_cluster_number_of_experienced_establisheds_array = 0;
        my $fake_dbsample_dates_cluster_number_of_naive_establisheds_array = 0;
        my $fake_dbsample_dates_cluster_number_of_notclassified_establisheds_array = 0;
        my $fake_dbsample_dates_cluster_number_of_drug_resistance_establisheds_array = 0;
        my $fake_dbsample_dates_cluster_number_of_experienced_drug_resistance_establisheds_array = 0;
        my $fake_dbsample_dates_cluster_number_of_naive_drug_resistance_establisheds_array = 0;
        my $fake_dbsample_dates_cluster_number_of_notclassified_drug_resistance_establisheds_array = 0;

        my @fake_random_dbsample_dates_temp_array = ();
        my @fake_random_dbsample_dates_recents_in_cluster_array = ();
        my @fake_random_dbsample_dates_cluster_number_of_recents_array = ();
        my @fake_random_dbsample_dates_cluster_number_of_experienced_recents_array = ();
        my @fake_random_dbsample_dates_cluster_number_of_naive_recents_array = ();
        my @fake_random_dbsample_dates_cluster_number_of_notclassified_recents_array = ();
        my @fake_random_dbsample_dates_cluster_number_of_drug_resistance_recents_array = ();
        my @fake_random_dbsample_dates_cluster_number_of_experienced_drug_resistance_recents_array = ();
        my @fake_random_dbsample_dates_cluster_number_of_naive_drug_resistance_recents_array = ();
        my @fake_random_dbsample_dates_cluster_number_of_notclassified_drug_resistance_recents_array = ();
        my @fake_random_dbsample_dates_cluster_number_of_establisheds_array = ();
        my @fake_random_dbsample_dates_cluster_number_of_experienced_establisheds_array = ();
        my @fake_random_dbsample_dates_cluster_number_of_naive_establisheds_array = ();
        my @fake_random_dbsample_dates_cluster_number_of_notclassified_establisheds_array = ();
        my @fake_random_dbsample_dates_cluster_number_of_drug_resistance_establisheds_array = ();
```

```
        my @fake_random_dbsample_dates_cluster_number_of_experienced_drug_resistance_establisheds_array =
();
        my @fake_random_dbsample_dates_cluster_number_of_naive_drug_resistance_establisheds_array = ();
        my @fake_random_dbsample_dates_cluster_number_of_notclassified_drug_resistance_establisheds_array
= ();
        my $fake_random_dbsample_dates_temp_array = 0;
        my $fake_random_dbsample_dates_recents_in_cluster_array = 0;
        my $fake_random_dbsample_dates_cluster_number_of_recents_array = 0;
        my $fake_random_dbsample_dates_cluster_number_of_experienced_recents_array = 0;
        my $fake_random_dbsample_dates_cluster_number_of_naive_recents_array = 0;
        my $fake_random_dbsample_dates_cluster_number_of_notclassified_recents_array = 0;
        my $fake_random_dbsample_dates_cluster_number_of_drug_resistance_recents_array = 0;
        my $fake_random_dbsample_dates_cluster_number_of_experienced_drug_resistance_recents_array = 0;
        my $fake_random_dbsample_dates_cluster_number_of_naive_drug_resistance_recents_array = 0;
        my $fake_random_dbsample_dates_cluster_number_of_notclassified_drug_resistance_recents_array = 0;
        my $fake_random_dbsample_dates_cluster_number_of_establisheds_array = 0;
        my $fake_random_dbsample_dates_cluster_number_of_experienced_establisheds_array = 0;
        my $fake_random_dbsample_dates_cluster_number_of_naive_establisheds_array = 0;
        my $fake_random_dbsample_dates_cluster_number_of_notclassified_establisheds_array = 0;
        my $fake_random_dbsample_dates_cluster_number_of_drug_resistance_establisheds_array = 0;
        my $fake_random_dbsample_dates_cluster_number_of_experienced_drug_resistance_establisheds_array =
0;
        my $fake_random_dbsample_dates_cluster_number_of_naive_drug_resistance_establisheds_array = 0;
        my $fake_random_dbsample_dates_cluster_number_of_notclassified_drug_resistance_establisheds_array
= 0;

        #Section to generate fake dbsample dated clusters
        my $count2 = 0;
        while($count2 < $iterations){
                #Must initially use the clusters stored in the recents hash, because these are a subset of
all clusters as determined by the size selection step when setting up the recents aray hash...
                foreach my $unique_cluster_id (sort keys %hash_of_recents_cluster_arrays){
                        my %dbsample_dates_hash = ();
                        my %fake_ordered_infector_hash = ();
                        my %fake_random_infector_hash = ();
                        #...but then, for these clusters, you must get all members (not just recents) to
set up your fake clusters
                        foreach my $sequence1 (@{$hash_of_cluster_arrays{$unique_cluster_id}}){
                                $dbsample_dates_hash{$sequence1} = $sequence_sampledate_hash{$sequence1};
                        }
                        #This variable covers the overall cluster timespan, not just recents
                        my @overall_cluster_members_sorted_by_dbsample_date = ();
                        @overall_cluster_members_sorted_by_dbsample_date = sort{$dbsample_dates_hash{$a}
cmp $dbsample_dates_hash{$b}} keys %dbsample_dates_hash;
                        #Section to assign random infection dates to infections to compare to scenarios
above
                        #Get timespan of cluster...
                        my $earliest_cluster_date =
$sequence_sampledate_hash{$overall_cluster_members_sorted_by_dbsample_date[0]};
                        my $latest_cluster_date =
$sequence_sampledate_hash{$overall_cluster_members_sorted_by_dbsample_date[@overall_cluster_members_sorted_
by_dbsample_date-1]};
                        my $cluster_timespan = $latest_cluster_date - $earliest_cluster_date;
                        #print
"earliest_cluster_date:$earliest_cluster_date;latest_cluster_date:$latest_cluster_date;timespan:$cluster_ti
mespan\n";
                        #Then assign each infection a new 'fake' dbsample date based on random selection
of date within timespan...
                        my %fake_dbsample_dates_hash = ();
                        my $this_count = 0;
                        #This loop goes through members of the original cluster in dbsample order until
the number matches the number of recents in the real cluster, these sequences, that have essentialy been
designated recent (but not randomly), then get assigned a random date of infection between the upper and
lower dbsample dates of the original cluster - thereby introducing the stochastic randoom element into the
model (i.e. essentially it doesn't matter what sequences are designated recent, as long as their infection
dates are randomised
                        foreach my $seq_id (@overall_cluster_members_sorted_by_dbsample_date){
                                if($this_count < scalar
@{$hash_of_recents_cluster_arrays{$unique_cluster_id}}){
                                        #This recent infection gets a random date assigned
                                        $fake_dbsample_dates_hash{$seq_id} = $earliest_cluster_date +
int(rand($cluster_timespan));
                                }
                                $this_count++;
                        }
                        my @fake_dbsample_dates_sorted_by_dbsample_date = ();
                        @fake_dbsample_dates_sorted_by_dbsample_date = sort{$fake_dbsample_dates_hash{$a}
cmp $fake_dbsample_dates_hash{$b}} keys %fake_dbsample_dates_hash;
                        my $array_count2 = 0;
                        foreach (@fake_dbsample_dates_sorted_by_dbsample_date){
                                if($array_count2 < scalar @fake_dbsample_dates_sorted_by_dbsample_date &&
$array_count2 > 0){

        $fake_ordered_infector_hash{$fake_dbsample_dates_sorted_by_dbsample_date[$array_count2]} =
$fake_dbsample_dates_sorted_by_dbsample_date[$array_count2 - 1];
                                }
                                $array_count2++;
                        }
```

```
                        @fake_dbsample_dates_temp_array =
@{ordersorter($unique_cluster_id,\%fake_ordered_infector_hash,\%fake_dbsample_dates_hash)};
                        push @fake_dbsample_dates_recents_in_cluster_array,
$fake_dbsample_dates_temp_array[0];
                        push @fake_dbsample_dates_cluster_number_of_recents_array,
$fake_dbsample_dates_temp_array[1];
                        push @fake_dbsample_dates_cluster_number_of_experienced_recents_array,
$fake_dbsample_dates_temp_array[2];
                        push @fake_dbsample_dates_cluster_number_of_naive_recents_array,
$fake_dbsample_dates_temp_array[3];
                        push @fake_dbsample_dates_cluster_number_of_notclassified_recents_array,
$fake_dbsample_dates_temp_array[4];
                        push @fake_dbsample_dates_cluster_number_of_drug_resistance_recents_array,
$fake_dbsample_dates_temp_array[5];
                        push
@fake_dbsample_dates_cluster_number_of_experienced_drug_resistance_recents_array,
$fake_dbsample_dates_temp_array[6];
                        push @fake_dbsample_dates_cluster_number_of_naive_drug_resistance_recents_array,
$fake_dbsample_dates_temp_array[7];
                        push
@fake_dbsample_dates_cluster_number_of_notclassified_drug_resistance_recents_array,
$fake_dbsample_dates_temp_array[8];
                        push @fake_dbsample_dates_cluster_number_of_establisheds_array,
$fake_dbsample_dates_temp_array[9];
                        push @fake_dbsample_dates_cluster_number_of_experienced_establisheds_array,
$fake_dbsample_dates_temp_array[10];
                        push @fake_dbsample_dates_cluster_number_of_naive_establisheds_array,
$fake_dbsample_dates_temp_array[11];
                        push @fake_dbsample_dates_cluster_number_of_notclassified_establisheds_array,
$fake_dbsample_dates_temp_array[12];
                        push @fake_dbsample_dates_cluster_number_of_drug_resistance_establisheds_array,
$fake_dbsample_dates_temp_array[13];
                        push
@fake_dbsample_dates_cluster_number_of_experienced_drug_resistance_establisheds_array,
$fake_dbsample_dates_temp_array[14];
                        push
@fake_dbsample_dates_cluster_number_of_naive_drug_resistance_establisheds_array,
$fake_dbsample_dates_temp_array[15];
                        push
@fake_dbsample_dates_cluster_number_of_notclassified_drug_resistance_establisheds_array,
$fake_dbsample_dates_temp_array[16];
                        $fake_dbsample_dates_recents_in_cluster_array =
$fake_dbsample_dates_recents_in_cluster_array + $fake_dbsample_dates_temp_array[0];
                        $fake_dbsample_dates_cluster_number_of_recents_array =
$fake_dbsample_dates_cluster_number_of_recents_array + $fake_dbsample_dates_temp_array[1];
                        $fake_dbsample_dates_cluster_number_of_experienced_recents_array =
$fake_dbsample_dates_cluster_number_of_experienced_recents_array + $fake_dbsample_dates_temp_array[2];
                        $fake_dbsample_dates_cluster_number_of_naive_recents_array =
$fake_dbsample_dates_cluster_number_of_naive_recents_array + $fake_dbsample_dates_temp_array[3];
                        $fake_dbsample_dates_cluster_number_of_notclassified_recents_array =
$fake_dbsample_dates_cluster_number_of_notclassified_recents_array + $fake_dbsample_dates_temp_array[4];
                        $fake_dbsample_dates_cluster_number_of_drug_resistance_recents_array =
$fake_dbsample_dates_cluster_number_of_drug_resistance_recents_array + $fake_dbsample_dates_temp_array[5];
                        $fake_dbsample_dates_cluster_number_of_experienced_drug_resistance_recents_array
= $fake_dbsample_dates_cluster_number_of_experienced_drug_resistance_recents_array +
$fake_dbsample_dates_temp_array[6];
                        $fake_dbsample_dates_cluster_number_of_naive_drug_resistance_recents_array =
$fake_dbsample_dates_cluster_number_of_naive_drug_resistance_recents_array +
$fake_dbsample_dates_temp_array[7];

        $fake_dbsample_dates_cluster_number_of_notclassified_drug_resistance_recents_array =
$fake_dbsample_dates_cluster_number_of_notclassified_drug_resistance_recents_array +
$fake_dbsample_dates_temp_array[8];
                        $fake_dbsample_dates_cluster_number_of_establisheds_array =
$fake_dbsample_dates_cluster_number_of_establisheds_array + $fake_dbsample_dates_temp_array[9];
                        $fake_dbsample_dates_cluster_number_of_experienced_establisheds_array =
$fake_dbsample_dates_cluster_number_of_experienced_establisheds_array +
$fake_dbsample_dates_temp_array[10];
                        $fake_dbsample_dates_cluster_number_of_naive_establisheds_array =
$fake_dbsample_dates_cluster_number_of_naive_establisheds_array + $fake_dbsample_dates_temp_array[11];
                        $fake_dbsample_dates_cluster_number_of_notclassified_establisheds_array =
$fake_dbsample_dates_cluster_number_of_notclassified_establisheds_array +
$fake_dbsample_dates_temp_array[12];
                        $fake_dbsample_dates_cluster_number_of_drug_resistance_establisheds_array =
$fake_dbsample_dates_cluster_number_of_drug_resistance_establisheds_array +
$fake_dbsample_dates_temp_array[13];

        $fake_dbsample_dates_cluster_number_of_experienced_drug_resistance_establisheds_array =
$fake_dbsample_dates_cluster_number_of_experienced_drug_resistance_establisheds_array +
$fake_dbsample_dates_temp_array[14];
                        $fake_dbsample_dates_cluster_number_of_naive_drug_resistance_establisheds_array =
$fake_dbsample_dates_cluster_number_of_naive_drug_resistance_establisheds_array +
$fake_dbsample_dates_temp_array[15];

        $fake_dbsample_dates_cluster_number_of_notclassified_drug_resistance_establisheds_array =
$fake_dbsample_dates_cluster_number_of_notclassified_drug_resistance_establisheds_array +
$fake_dbsample_dates_temp_array[16];
```

```
                        %fake_random_infector_hash =
random_infector_assignment(\@fake_dbsample_dates_sorted_by_dbsample_date, \%fake_dbsample_dates_hash,
$unique_cluster_id);
                        @fake_random_dbsample_dates_temp_array =
@{ordersorter($unique_cluster_id,\%fake_random_infector_hash,\%fake_dbsample_dates_hash)};
                        push @fake_random_dbsample_dates_recents_in_cluster_array,
$fake_random_dbsample_dates_temp_array[0];
                        push @fake_random_dbsample_dates_cluster_number_of_recents_array,
$fake_random_dbsample_dates_temp_array[1];
                        push @fake_random_dbsample_dates_cluster_number_of_experienced_recents_array,
$fake_random_dbsample_dates_temp_array[2];
                        push @fake_random_dbsample_dates_cluster_number_of_naive_recents_array,
$fake_random_dbsample_dates_temp_array[3];
                        push @fake_random_dbsample_dates_cluster_number_of_notclassified_recents_array,
$fake_random_dbsample_dates_temp_array[4];
                        push @fake_random_dbsample_dates_cluster_number_of_drug_resistance_recents_array,
$fake_random_dbsample_dates_temp_array[5];
                        push
@fake_random_dbsample_dates_cluster_number_of_experienced_drug_resistance_recents_array,
$fake_random_dbsample_dates_temp_array[6];
                        push
@fake_random_dbsample_dates_cluster_number_of_naive_drug_resistance_recents_array,
$fake_random_dbsample_dates_temp_array[7];
                        push
@fake_random_dbsample_dates_cluster_number_of_notclassified_drug_resistance_recents_array,
$fake_random_dbsample_dates_temp_array[8];
                        push @fake_random_dbsample_dates_cluster_number_of_establisheds_array,
$fake_random_dbsample_dates_temp_array[9];
                        push
@fake_random_dbsample_dates_cluster_number_of_experienced_establisheds_array,
$fake_random_dbsample_dates_temp_array[10];
                        push @fake_random_dbsample_dates_cluster_number_of_naive_establisheds_array,
$fake_random_dbsample_dates_temp_array[11];
                        push
@fake_random_dbsample_dates_cluster_number_of_notclassified_establisheds_array,
$fake_random_dbsample_dates_temp_array[12];
                        push
@fake_random_dbsample_dates_cluster_number_of_drug_resistance_establisheds_array,
$fake_random_dbsample_dates_temp_array[13];
                        push
@fake_random_dbsample_dates_cluster_number_of_experienced_drug_resistance_establisheds_array,
$fake_random_dbsample_dates_temp_array[14];
                        push
@fake_random_dbsample_dates_cluster_number_of_naive_drug_resistance_establisheds_array,
$fake_random_dbsample_dates_temp_array[15];
                        push
@fake_random_dbsample_dates_cluster_number_of_notclassified_drug_resistance_establisheds_array,
$fake_random_dbsample_dates_temp_array[16];
                        $fake_random_dbsample_dates_recents_in_cluster_array =
$fake_random_dbsample_dates_recents_in_cluster_array + $fake_random_dbsample_dates_temp_array[0];
                        $fake_random_dbsample_dates_cluster_number_of_recents_array =
$fake_random_dbsample_dates_cluster_number_of_recents_array + $fake_random_dbsample_dates_temp_array[1];
                        $fake_random_dbsample_dates_cluster_number_of_experienced_recents_array =
$fake_random_dbsample_dates_cluster_number_of_experienced_recents_array +
$fake_random_dbsample_dates_temp_array[2];
                        $fake_random_dbsample_dates_cluster_number_of_naive_recents_array =
$fake_random_dbsample_dates_cluster_number_of_naive_recents_array +
$fake_random_dbsample_dates_temp_array[3];
                        $fake_random_dbsample_dates_cluster_number_of_notclassified_recents_array =
$fake_random_dbsample_dates_cluster_number_of_notclassified_recents_array +
$fake_random_dbsample_dates_temp_array[4];
                        $fake_random_dbsample_dates_cluster_number_of_drug_resistance_recents_array =
$fake_random_dbsample_dates_cluster_number_of_drug_resistance_recents_array +
$fake_random_dbsample_dates_temp_array[5];

        $fake_random_dbsample_dates_cluster_number_of_experienced_drug_resistance_recents_array =
$fake_random_dbsample_dates_cluster_number_of_experienced_drug_resistance_recents_array +
$fake_random_dbsample_dates_temp_array[6];
                        $fake_random_dbsample_dates_cluster_number_of_naive_drug_resistance_recents_array
= $fake_random_dbsample_dates_cluster_number_of_naive_drug_resistance_recents_array +
$fake_random_dbsample_dates_temp_array[7];

        $fake_random_dbsample_dates_cluster_number_of_notclassified_drug_resistance_recents_array =
$fake_random_dbsample_dates_cluster_number_of_notclassified_drug_resistance_recents_array +
$fake_random_dbsample_dates_temp_array[8];
                        $fake_random_dbsample_dates_cluster_number_of_establisheds_array =
$fake_random_dbsample_dates_cluster_number_of_establisheds_array +
$fake_random_dbsample_dates_temp_array[9];
                        $fake_random_dbsample_dates_cluster_number_of_experienced_establisheds_array =
$fake_random_dbsample_dates_cluster_number_of_experienced_establisheds_array +
$fake_random_dbsample_dates_temp_array[10];
                        $fake_random_dbsample_dates_cluster_number_of_naive_establisheds_array =
$fake_random_dbsample_dates_cluster_number_of_naive_establisheds_array +
$fake_random_dbsample_dates_temp_array[11];
                        $fake_random_dbsample_dates_cluster_number_of_notclassified_establisheds_array =
$fake_random_dbsample_dates_cluster_number_of_notclassified_establisheds_array +
$fake_random_dbsample_dates_temp_array[12];
```

```
                        $fake_random_dbsample_dates_cluster_number_of_drug_resistance_establisheds_array
= $fake_random_dbsample_dates_cluster_number_of_drug_resistance_establisheds_array +
$fake_random_dbsample_dates_temp_array[13];

        $fake_random_dbsample_dates_cluster_number_of_experienced_drug_resistance_establisheds_array =
$fake_random_dbsample_dates_cluster_number_of_experienced_drug_resistance_establisheds_array +
$fake_random_dbsample_dates_temp_array[14];

        $fake_random_dbsample_dates_cluster_number_of_naive_drug_resistance_establisheds_array =
$fake_random_dbsample_dates_cluster_number_of_naive_drug_resistance_establisheds_array +
$fake_random_dbsample_dates_temp_array[15];

        $fake_random_dbsample_dates_cluster_number_of_notclassified_drug_resistance_establisheds_array =
$fake_random_dbsample_dates_cluster_number_of_notclassified_drug_resistance_establisheds_array +
$fake_random_dbsample_dates_temp_array[16];
                }
                $count2++;
        }

        $fake1 = $fake_dbsample_dates_recents_in_cluster_array/$iterations;
        print "fake:total_of_recent_infection_events_in_clusters:$fake1;";
        $fake2 = $fake_dbsample_dates_cluster_number_of_recents_array/$iterations;
        print "fake:total_of_recent_phase_infections:$fake2;";

        my $temp_fake = $fake1-$fake2;

        system ("rm -rf fisher.rscript.R");
        system ("rm -rf fisher.Rout.txt");
        system ("rm -rf fisher.rscript.Rout");

        my $temp_random = $random1-$random2;
        open(RSCRIPT, ">fisher.rscript.R");
        print RSCRIPT "x <- matrix(c($random2,$temp_random,$fake2,$temp_fake),nc=2)\n";
        print RSCRIPT "y <- fisher.test(x)\n";
        print RSCRIPT "write.table(print(c(y\$p.value,y\$estimate,y\$conf.int)), file='fisher.Rout.txt',
quote=F, row.names=F, col.names=F, eol='\\n')\n";
        close RSCRIPT;

        system ("R CMD BATCH fisher.rscript.R");

        open(RSCRIPTINPUT, "fisher.Rout.txt");

        my $counter1 = 0;
        my @fisher_array1;
        while(my $fileline = <RSCRIPTINPUT>){
                chomp($fileline);
                $fisher_array1[$counter1] = $fileline;
                $counter1++;
        }
        print "random_vs_fake;p-
value;$fisher_array1[0];odds_ratio;$fisher_array1[1];95pc_lower_conf_int;$fisher_array1[2];95pc_upper_conf_
int;$fisher_array1[3];";

        system ("rm -rf fisher.rscript.R");
        system ("rm -rf fisher.Rout.txt");
        system ("rm -rf fisher.rscript.Rout");

        $fake_random1 = $fake_random_dbsample_dates_recents_in_cluster_array/$iterations;
        print "fake_random:total_of_recent_infection_events_in_clusters:$fake_random1;";
        $fake_random2 = $fake_random_dbsample_dates_cluster_number_of_recents_array/$iterations;
        print "fake_random:total_of_recent_phase_infections:$fake_random2;";

        my $temp_fake_random = $fake_random1-$fake_random2;
        open(RSCRIPT, ">fisher.rscript.R");
        print RSCRIPT "x <- matrix(c($random2,$temp_random,$fake_random2,$temp_fake_random),nc=2)\n";
        print RSCRIPT "y <- fisher.test(x)\n";
        print RSCRIPT "write.table(print(c(y\$p.value,y\$estimate,y\$conf.int)), file='fisher.Rout.txt',
quote=F, row.names=F, col.names=F, eol='\\n')\n";
        close RSCRIPT;

        system ("R CMD BATCH fisher.rscript.R");

        open(RSCRIPTINPUT, "fisher.Rout.txt");

        my $counter2 = 0;
        my @fisher_array2;
        while(my $fileline = <RSCRIPTINPUT>){
                chomp($fileline);
                $fisher_array2[$counter2] = $fileline;
                $counter2++;
        }
        print "random_vs_fake_random;p-
value;$fisher_array2[0];odds_ratio;$fisher_array2[1];95pc_lower_conf_int;$fisher_array2[2];95pc_upper_conf_
int;$fisher_array2[3];";

        system ("rm -rf fisher.rscript.R");
        system ("rm -rf fisher.Rout.txt");
        system ("rm -rf fisher.rscript.Rout");
```

```
        my $temp_ordered = $ordered1-$ordered2;
        open(RSCRIPT, ">fisher.rscript.R");
        print RSCRIPT "x <- matrix(c($ordered2,$temp_ordered,$fake2,$temp_fake),nc=2)\n";
        print RSCRIPT "y <- fisher.test(x)\n";
        print RSCRIPT "write.table(print(c(y\$p.value,y\$estimate,y\$conf.int)), file='fisher.Rout.txt',
quote=F, row.names=F, col.names=F, eol='\\n')\n";
        close RSCRIPT;

        system ("R CMD BATCH fisher.rscript.R");

        open(RSCRIPTINPUT, "fisher.Rout.txt");

        my $counter3 = 0;
        my @fisher_array3;
        while(my $fileline = <RSCRIPTINPUT>){
                chomp($fileline);
                $fisher_array3[$counter3] = $fileline;
                $counter3++;
        }
        print "ordered_vs_fake;p-
value;$fisher_array3[0];odds_ratio;$fisher_array3[1];95pc_lower_conf_int;$fisher_array3[2];95pc_upper_conf_
int;$fisher_array3[3];";

}else{
        print "#!!# No clusters to process given the current criteria for numbers of recents #!!#;";
}

print "iterations:$iterations;";

print "number_of_suitable_clusters:$number_of_suitable_clusters\n";

close INPUT1;
close INPUT2;
close INPUT3;
#close OUTPUT;

sub random_infector_assignment{
        my @same_first_date_array = ();
        my ($cluster_members_sorted_by_dbsample_date_ref, $dbsample_dates_hash_ref, $unique_cluster_id) =
@_;
        my @cluster_members_sorted_by_dbsample_date = @{$cluster_members_sorted_by_dbsample_date_ref};
        my %dbsample_dates_hash = %{$dbsample_dates_hash_ref};
        foreach my $temp_key (@cluster_members_sorted_by_dbsample_date){
                if($temp_key != $cluster_members_sorted_by_dbsample_date[0] &&
$dbsample_dates_hash{$temp_key} == $dbsample_dates_hash{$cluster_members_sorted_by_dbsample_date[0]}){
                        push @same_first_date_array, $temp_key;
                }
        }
        ##Random assignment section
        my %random_infector_hash = ();
        foreach my $sequence1 (@cluster_members_sorted_by_dbsample_date){
                my @potential_infector_array;
                foreach my $sequence2 (@cluster_members_sorted_by_dbsample_date){
                        if($sequence1 != $sequence2 && $sequence_sampledate_hash{$sequence2} <
$sequence_sampledate_hash{$sequence1}){
                                push @potential_infector_array, $sequence2;
                        }
                }
                my $random_array_element;
                if(scalar @potential_infector_array > 0){
                        $random_infector_hash{$sequence1} = $potential_infector_array[rand
@potential_infector_array];
                }
        }
        #Check you have two or more seqs with same date that is the oldest date in the cluster recents, if
so then make the original one from above ordering the random infector of the others, as they all have same
date, sothen just add the first sequence with the same first date to all the subsequent ones, to make sure
you have them all
        if(scalar @same_first_date_array > 0){
                #first add in the original oldest sequence from above, so you have all of them
                push @same_first_date_array, $cluster_members_sorted_by_dbsample_date[0];
                my @shuffled_same_first_date_array = shuffle(@same_first_date_array);
                my $array_count = 0;
                foreach (@shuffled_same_first_date_array){
                        if($array_count < scalar @shuffled_same_first_date_array && $array_count > 0){
                                $random_infector_hash{$shuffled_same_first_date_array[$array_count]} =
$shuffled_same_first_date_array[$array_count - 1];
                        }
                        $array_count++;
                }
        }
        return %random_infector_hash;
}

sub ordersorter{
        my ($unique_cluster_id, $infector_hash_ref, $temp_dbsample_date_hash_ref) = @_;
        my %infector_hash = %{$infector_hash_ref};
```

```
        my %temp_dbsample_date_hash = %{$temp_dbsample_date_hash_ref};
        my $recents_in_cluster = scalar keys %infector_hash;
        my $recent = 0;
        my $experienced_recent = 0;
        my $naive_recent = 0;
        my $notclassified_recent = 0;
        my $drug_resistance_recent = 0;
        my $experienced_drug_resistance_recent = 0;
        my $naive_drug_resistance_recent = 0;
        my $notclassified_drug_resistance_recent = 0;

        my $established = 0;
        my $experienced_established = 0;
        my $naive_established = 0;
        my $notclassified_established = 0;
        my $drug_resistance_established = 0;
        my $experienced_drug_resistance_established = 0;
        my $naive_drug_resistance_established = 0;
        my $notclassified_drug_resistance_established = 0;

        foreach my $sequence1 (sort keys %infector_hash){
                #print "$sequence1 $temp_dbsample_date_hash{$sequence1} $infector_hash{$sequence1}
$temp_dbsample_date_hash{$infector_hash{$sequence1}}\n";
                # This foreach goes through each sequence, and then the inner while compares each outer
sequence to all the other cluster members
                #$recents_in_cluster++;
                if($temp_dbsample_date_hash{$sequence1} - $infection_window_for_recents <
$temp_dbsample_date_hash{$infector_hash{$sequence1}}){
                        $recent++;
                        if($sequence_therapy_status_hash{$infector_hash{$sequence1}} eq '"Experienced"'){
                                $experienced_recent++;
                                if($sequence_drug_resistance_status_hash{$infector_hash{$sequence1}} eq
"D"){
                                        $drug_resistance_recent++;
                                        $experienced_drug_resistance_recent++;
                                }
                        }elsif($sequence_therapy_status_hash{$infector_hash{$sequence1}} eq '"Naive"'){
                                $naive_recent++;
                                if($sequence_drug_resistance_status_hash{$infector_hash{$sequence1}} eq
"D"){
                                        $drug_resistance_recent++;
                                        $naive_drug_resistance_recent++;
                                }
                        }elsif($sequence_therapy_status_hash{$infector_hash{$sequence1}} eq '"Not
classified"'){
                                $notclassified_recent++;
                                if($sequence_drug_resistance_status_hash{$infector_hash{$sequence1}} eq
"D"){
                                        $drug_resistance_recent++;
                                        $notclassified_drug_resistance_recent++;
                                }
                        }
                }else{
                        $established++;
                        if($sequence_therapy_status_hash{$infector_hash{$sequence1}} eq '"Experienced"'){
                                $experienced_established++;
                                if($sequence_drug_resistance_status_hash{$infector_hash{$sequence1}} eq
"D"){
                                        $drug_resistance_established++;
                                        $experienced_drug_resistance_established++;
                                }
                        }elsif($sequence_therapy_status_hash{$infector_hash{$sequence1}} eq '"Naive"'){
                                $naive_established++;
                                if($sequence_drug_resistance_status_hash{$infector_hash{$sequence1}} eq
"D"){
                                        $drug_resistance_established++;
                                        $naive_drug_resistance_established++;
                                }
                        }elsif($sequence_therapy_status_hash{$infector_hash{$sequence1}} eq '"Not
classified"'){
                                $notclassified_established++;
                                if($sequence_drug_resistance_status_hash{$infector_hash{$sequence1}} eq
"D"){
                                        $drug_resistance_established++;
                                $notclassified_drug_resistance_established++;
                                }
                        }
                }
        }

        my @array_of_results = ($recents_in_cluster, $recent, $experienced_recent, $naive_recent,
$notclassified_recent, $drug_resistance_recent, $experienced_drug_resistance_recent,
$naive_drug_resistance_recent, $notclassified_drug_resistance_recent, $established,
$experienced_established, $naive_established, $notclassified_established, $drug_resistance_established,
$experienced_drug_resistance_established, $naive_drug_resistance_established,
$notclassified_drug_resistance_established);
        return \@array_of_results;
}
```

genetic_distance_model.pl

```
#!usr/bin/perl -w

use strict;
use warnings;

#
# File to find number of recent and established infections within a cluster that could have seeded other
cluster infections
#

if (defined($ARGV[0])) {
  open(INPUT1, $ARGV[0]);
} else {
  print 'No recent/established status file of the form "sequenceId sampleDate recent/establishedStatus"\n';
        my $input = <>;
  open(INPUT1, $input);
}

if (defined($ARGV[1])) {
  open(INPUT2, $ARGV[1]);
} else {
  print "No treatment experience file, please enter one:\n";
        my $input = <>;
  open(INPUT2, $input);
}

if (defined($ARGV[2])) {
  open(INPUT3, $ARGV[2]);

} else {
  print "No clusters file, please enter one:\n";
        my $input = <>;
  open(INPUT3, $input);
}

#Input4 is a file that has been preprocessed by an awk command such as: awk '($9 !~ /NA/ && $9 <= 0.03) &&
$0 !~ /^mean/'
/home/virology/Perl/re_doing/REGA_subtype_subsets/DRAM_subsets/findcluster94.5/REGA_b_subtype_subset_non_du
plicate_DRAM_sequences.94.5.0.2000.findclusteroutput.txt  > testing_short_distances
#In other words, it is a list of pairwise distances generated by the findclusterdistances.pl script - but
this can be easily adapted
if (defined($ARGV[3])) {
  open(INPUT4, $ARGV[3]);
} else {
  print "No pairwise distances file, please enter one:\n";
        my $input = <>;
  open(INPUT4, $input);
}
#my $output = "genetic_distance_model_output";
#open(OUTPUT, ">$output");

my %sequence_infection_status_hash = ();
my %sequence_sampledate_hash = ();
my %sequence_therapy_status_hash = ();
my %sequence_drug_resistance_status_hash = ();
my %hash_of_cluster_arrays = ();
my %hash_of_cluster_lengths = ();
my %hash_of_pairwise_distances = ();
my %hash_of_recent_phase = ();
my $infection_window_for_recents = 125;
while(my $fileline = <INPUT1>){
        chomp($fileline);
        # INPUT1 should be a tab delimited list of sequence id followed by dbsample date as a number
followed by recent or established status followed by drug resistance (D for drug resistance, S for drug
sensitive)
        my @fileline = split(/\t/,$fileline);
        # Use sequence id as key, and status as value
        if(!$sequence_infection_status_hash{$fileline[0]}){
                $sequence_sampledate_hash{$fileline[0]} = $fileline[1];
                $sequence_infection_status_hash{$fileline[0]} = $fileline[2];
                $sequence_drug_resistance_status_hash{$fileline[0]} = $fileline[3];
        }else{
                die "Sequence $fileline[0] used more than once!\n";
        }
}
while(my $fileline = <INPUT2>){
        chomp($fileline);
        # INPUT2 should be a tab delimited list of sequence id followed by therapy status
        my @fileline = split(/,/,$fileline);
        # Use sequence id as key, and status as value
        if(!$sequence_therapy_status_hash{$fileline[0]}){
                $sequence_therapy_status_hash{$fileline[0]} = $fileline[1];
        }else{
                die "Sequence $fileline[0] used more than once!\n";
```

```
        }
}
while(my $fileline = <INPUT3>){
        if($fileline =~ /^\s\d/){
                chomp($fileline);
                # USE // WITH split() NOT '' with this form of cluster id - using // means first element
is empty, because there's a space at the beginning of the line e.g. " 101253 112376 113698"
                my @clusterarray = split(/\s/, $fileline);
                splice(@clusterarray,0,1);
                $hash_of_cluster_arrays{$fileline} = [@clusterarray];
                $hash_of_cluster_lengths{$fileline} = @clusterarray;
        }
}
while(my $fileline = <INPUT4>){
        chomp($fileline);
        if($fileline =~ /Distance between node (\d+) and node (\d+) is (\d+\.\d+)/){
                #This just generates a unique key of the form largestSequenceId_smallestSequenceId
                if($1 > $2){
                        my $key = $1."_".$2;
                        $hash_of_pairwise_distances{$key} = $3;
                }else{
                        my $key = $2."_".$1;
                        $hash_of_pairwise_distances{$key} = $3;
                }
        }
}
foreach my $unique_cluster_id (sort keys %hash_of_cluster_arrays){
        my %hash_of_pairwise_distances_done = ();
        my %hash_to_check_overcount = ();
        my @temp_array = ();
        foreach my $sequence1 (@{$hash_of_cluster_arrays{$unique_cluster_id}}){
                foreach my $sequence2 (@{$hash_of_cluster_arrays{$unique_cluster_id}}){
                        my $key = 0;
                        if($sequence1 != $sequence2){
                                if($sequence1 > $sequence2){
                                        $key = $sequence1."_".$sequence2;
                                }else{
                                        $key = $sequence2."_".$sequence1;
                                }
                        }
                        if($hash_of_pairwise_distances{$key} && !$hash_of_pairwise_distances_done{$key}){
                                $hash_of_pairwise_distances_done{$key} = 1;
                                my @temp_nodes_array = split(/_/,$key);

                                if(($sequence_infection_status_hash{$temp_nodes_array[0]} eq "recent") &&
($sequence_infection_status_hash{$temp_nodes_array[1]} eq "recent")){
                                        if(abs($sequence_sampledate_hash{$temp_nodes_array[0]} -
$sequence_sampledate_hash{$temp_nodes_array[1]}) <= $infection_window_for_recents){
                                                #This checks that you don't overcount transmission
because of something like A-B, A-C and B-C: one of these is excluded by the other two (if you do not count
super-infection!)
                                                if($hash_to_check_overcount{$temp_nodes_array[0]} &&
$hash_to_check_overcount{$temp_nodes_array[1]}){
                                                        print "overcount: $temp_nodes_array[0],
$temp_nodes_array[1]\n";
                                                }else{
                                                        $hash_to_check_overcount{$temp_nodes_array[0]}
= 1;
                                                        $hash_to_check_overcount{$temp_nodes_array[1]}
= 1;
                                                        push @temp_array,
$temp_nodes_array[0]."_".$temp_nodes_array[1];
                                                }
                                        }
                                }
                        }
                }
        }
        if(scalar(@temp_array) > 0){
                $hash_of_recent_phase{$unique_cluster_id} = \@temp_array;
        }
}

foreach my $unique_cluster_id (sort keys %hash_of_recent_phase){
        print "\n$unique_cluster_id\n";
        my @array = @{$hash_of_recent_phase{$unique_cluster_id}};
        foreach my $pair (@array){
                print "$pair\n";
        }
}

close INPUT1;
close INPUT2;
close INPUT3;
close INPUT4;
#close OUTPUT;
```

nucleotide_type_counter.pl

```
#!usr/bin/perl -w

# Program for taking sequences and converting them into a IUPAC containing consensus sequence
use Bio::SimpleAlign;
use Bio::AlignIO;
use Bio::SeqIO;
use Bio::PopGen::Utilities;
use Bio::Tools::Run::StandAloneBlast;
use Data::Dumper;

if (defined($ARGV[0])) {$sequences_input_file = $ARGV[0]} else {die "Missing file";}
if (defined($ARGV[1])) {$sanger_pol_seqs_file = $ARGV[1]} else {die "Missing file";}

my $sample_name = $sequences_input_file;
$sample_name =~ s/_usearch_primerstripped_pol_f_and_r_main_orf390_aligned_degapped\.fasta3//g;

my $aln_obj = Bio::AlignIO->new(-file => $sequences_input_file);
my $seq_obj = Bio::SeqIO->new(-file => $sanger_pol_seqs_file, -format => 'fasta');

my %all_sanger_pols_hash = ();
my $pol_sanger_seq_obj;
my $temp_sample_name = $sample_name."_sanger";
while (my $seq = $seq_obj->next_seq()){
        if($seq->display_id eq ($temp_sample_name)){
                $pol_sanger_seq_obj = $seq;
        }
}

my $aln = $aln_obj->next_aln();

my $pop = Bio::PopGen::Utilities->aln_to_population(-alignment=>$aln, -include_monomorphic => 1);
my @marker_names = $pop->get_marker_names;
my %predominant_variant_hash = ();
foreach $marker_name (@marker_names){
        my $variant_count = 0;
        my %temp_allele_hash = ();
#       print "marker_name:$marker_name\t";
        my $marker = $pop->get_Marker($marker_name);
        my %allele_freqs = $marker->get_Allele_Frequencies;
        foreach my $allele (keys %allele_freqs){
                #get any allele above a minimum threshold and store it in a temporary hash
                if($allele_freqs{$allele} >= 0.02){
                        $variant_count++;
                        $temp_allele_hash{$allele} = $allele_freqs{$allele};
                }
        }
        my $switch = 0;
        my $position = 0;
        foreach my $allele_ordered (sort {$temp_allele_hash{$b} <=> $temp_allele_hash{$a}} (keys
%temp_allele_hash)){
#                print "$allele_ordered\t$temp_allele_hash{$allele_ordered}\t";
                #this switch is just to generate a 'consensus' sequence with the most predomiant allele at
each position
                if(!$switch){
                        $position = $marker_name;
                        #convert marker name to number by removing prefix
                        $position =~ s/Site-//g;
                        $predominant_variant_hash{$position} = $allele_ordered;
                        $switch = 1;
                }
        }
#       print "\n";
}
my $counter = 0;
my $predominant_variant;
foreach my $key (sort {$a <=> $b} (keys %predominant_variant_hash)){
        $predominant_variant .= $predominant_variant_hash{$key};
}

my $predominant_variant_seq_obj = new Bio::Seq->new(-display_id => $sample_name, -seq =>
$predominant_variant);

my $blast_factory = Bio::Tools::Run::StandAloneBlast->new(-outfile => 'bl2seq.out', -program => "blastn");
my $blast_report = $blast_factory->bl2seq($predominant_variant_seq_obj,$pol_sanger_seq_obj);
$align_obj = Bio::AlignIO->new(-file => 'bl2seq.out', -format => 'bl2seq');

my $hsp = $blast_report->next_result->next_hit->next_hsp;
#print Dumper($hsp);
my $predom_start = $hsp->{'QUERY_START'};
my $pol_start = $hsp->{'HIT_START'};

my $pairwise_alignment = $align_obj->next_aln();

my $mapping_from_unaligned_predominant_variant_seq_to_aligned_predominant_variant_seq = $predom_start - 1;
```

```
my @reduced_marker_names;
#correct a copy of the original marker names with the mapping info
my $index = $mapping_from_unaligned_predominant_variant_seq_to_aligned_predominant_variant_seq;
#print "index:$index\n";
my @copy_marker_names = @marker_names;
@reduced_marker_names = splice(@copy_marker_names, $index);

my @aligned_seq_array = ();
my $array_count = 0;
foreach my $aligned_seq ($pairwise_alignment->each_seq()){
        if($aligned_seq->display_id =~ /sanger/){
                $pol_sanger_seq_aligned = $aligned_seq;
        }else{
                $predominant_variant_aligned = $aligned_seq;
        }
}

#print $predominant_variant_aligned->seq."\n";
#print $pol_sanger_seq_aligned->seq."\n";

my @predominant_variant_aligned_array = split("", $predominant_variant_aligned->seq);
my @pol_sanger_seq_aligned = split("", $pol_sanger_seq_aligned->seq);

$length = scalar @predominant_variant_aligned_array;
my $predom_gaps = 0;
my $acgt_agreements = 0;
my $iupac_agreements = 0;
my $disagreements = 0;
my $second_most_abundant_variant_agreements = 0;
my $gap_disagreements = 0;
for($i=0; $i<$length; $i++){
        if($predominant_variant_aligned_array[$i] eq $pol_sanger_seq_aligned[$i]){
#               print "agree:\t$predominant_variant_aligned_array[$i]\t$pol_sanger_seq_aligned[$i]\n";
                $acgt_agreements++;
        }else{

                $disagreements++;
                if($predominant_variant_aligned_array[$i] eq "-"){
                        $gap_disagreements++;
                        $predom_gaps++;
                }else{
                        my $temp_marker_name = $reduced_marker_names[$i-$predom_gaps];
                        my $temp_marker = $pop->get_Marker($temp_marker_name);
                        my %temp_allele_freqs = $temp_marker->get_Allele_Frequencies;
                        my @temp_alleles_array = ();
                        my $temp_alleles_string = ();
                        my $temp_variant_count = 0;
                        foreach my $temp_allele (sort {$temp_allele_freqs{$b} <=> $temp_allele_freqs{$a}}
(keys %temp_allele_freqs)){
                                #get any allele above a minimum threshold and store it in a temporary
hash - this cut-off can be something to do with when mixed base peaks start appearing, e.g.20-30% - but
actually I have seen that there is better iupac agreement with lower cut-offs, e.g. there are a few
instances where the ~11% minority variant still appears to contribute to the ambiguity in sanger (e.g.
Pool1_TC10)
                                print
"disag$i:\t$predominant_variant_aligned_array[$i]\t$pol_sanger_seq_aligned[$i]\t$temp_allele
$temp_allele_freqs{$temp_allele}\n";
                                        if($temp_allele_freqs{$temp_allele} >= 0.02){
#                                       print "\t$temp_allele $temp_allele_freqs{$temp_allele}";
                                        $temp_allele =~ tr/A-Z/a-z/;
                                        push(@temp_alleles_array, $temp_allele);
                                        $temp_alleles_string .= $temp_allele;
                                        $temp_variant_count++;
                                }
                        }
                        $temp_alleles_string =~ tr/uU/tT/;
                        if($temp_alleles_string =~ /a/ && $temp_alleles_string =~ /g/ &&
$temp_variant_count == 2){
                                if($pol_sanger_seq_aligned[$i] eq "r"){
#                                       print "\tiupac_agrees:\tr\t$pol_sanger_seq_aligned[$i]";
                                        $iupac_agreements++;
                                }else{
#                                       print "\tiupac_disagrees:\tr\t$pol_sanger_seq_aligned[$i]";
                                        if($temp_alleles_array[1] eq $pol_sanger_seq_aligned[$i]){
                                                $second_most_abundant_variant_agreements++;
                                        }
                                }
                        }elsif($temp_alleles_string =~ /c/ && $temp_alleles_string =~ /t/ &&
$temp_variant_count == 2){
                                if($pol_sanger_seq_aligned[$i] eq "y"){
#                                       print "\tiupac_agrees:\ty\t$pol_sanger_seq_aligned[$i]";
                                        $iupac_agreements++;
                                }else{
#                                       print "\tiupac_disagrees:\ty\t$pol_sanger_seq_aligned[$i]";
                                        if($temp_alleles_array[1] eq $pol_sanger_seq_aligned[$i]){
                                                $second_most_abundant_variant_agreements++;
                                        }
```

```
                        }
                    }elsif($temp_alleles_string =~ /c/ && $temp_alleles_string =~ /a/ &&
$temp_variant_count == 2){
                        if($pol_sanger_seq_aligned[$i] eq "m"){
#                           print "\tiupac_agrees:\tm\t$pol_sanger_seq_aligned[$i]";
                            $iupac_agreements++;
                        }else{
#                           print "\tiupac_disagrees:\tm\t$pol_sanger_seq_aligned[$i]";
                            if($temp_alleles_array[1] eq $pol_sanger_seq_aligned[$i]){
                                $second_most_abundant_variant_agreements++;
                            }
                        }
                    }elsif($temp_alleles_string =~ /t/ && $temp_alleles_string =~ /g/ &&
$temp_variant_count == 2){
                        if($pol_sanger_seq_aligned[$i] eq "k"){
#                           print "\tiupac_agrees:\tk\t$pol_sanger_seq_aligned[$i]";
                            $iupac_agreements++;
                        }else{
#                           print "\tiupac_disagrees:\tk\t$pol_sanger_seq_aligned[$i]";
                            if($temp_alleles_array[1] eq $pol_sanger_seq_aligned[$i]){
                                $second_most_abundant_variant_agreements++;
                            }
                        }
                    }elsif($temp_alleles_string =~ /t/ && $temp_alleles_string =~ /a/ &&
$temp_variant_count == 2){
                        if($pol_sanger_seq_aligned[$i] eq "w"){
#                           print "\tiupac_agrees:\tw\t$pol_sanger_seq_aligned[$i]";
                            $iupac_agreements++;
                        }else{
#                           print "\tiupac_disagrees:\tw\t$pol_sanger_seq_aligned[$i]";
                            if($temp_alleles_array[1] eq $pol_sanger_seq_aligned[$i]){
                                $second_most_abundant_variant_agreements++;
                            }
                        }
                    }elsif($temp_alleles_string =~ /c/ && $temp_alleles_string =~ /g/ &&
$temp_variant_count == 2){
                        if($pol_sanger_seq_aligned[$i] eq "s"){
#                           print "\tiupac_agrees:\ts\t$pol_sanger_seq_aligned[$i]";
                            $iupac_agreements++;
                        }else{
#                           print "\tiupac_disagrees:\ts\t$pol_sanger_seq_aligned[$i]";
                            if($temp_alleles_array[1] eq $pol_sanger_seq_aligned[$i]){
                                $second_most_abundant_variant_agreements++;
                            }
                        }
                    }elsif($temp_alleles_string =~ /c/ && $temp_alleles_string =~ /t/ &&
$temp_alleles_string =~ /g/ && $temp_variant_count == 3){
                        if($pol_sanger_seq_aligned[$i] eq "b"){
#                           print "\tiupac_agrees:\tb\t$pol_sanger_seq_aligned[$i]";
                            $iupac_agreements++;
                        }else{
#                           print "\tiupac_disagrees:\tb\t$pol_sanger_seq_aligned[$i]";
                            if($temp_alleles_array[1] eq $pol_sanger_seq_aligned[$i]){
                                $second_most_abundant_variant_agreements++;
                            }
                        }
                    }elsif($temp_alleles_string =~ /a/ && $temp_alleles_string =~ /t/ &&
$temp_alleles_string =~ /g/ && $temp_variant_count == 3){
                        if($pol_sanger_seq_aligned[$i] eq "d"){
#                           print "\tiupac_agrees:\td\t$pol_sanger_seq_aligned[$i]";
                            $iupac_agreements++;
                        }else{
#                           print "\tiupac_disagrees:\td\t$pol_sanger_seq_aligned[$i]";
                            if($temp_alleles_array[1] eq $pol_sanger_seq_aligned[$i]){
                                $second_most_abundant_variant_agreements++;
                            }
                        }
                    }elsif($temp_alleles_string =~ /a/ && $temp_alleles_string =~ /t/ &&
$temp_alleles_string =~ /c/ && $temp_variant_count == 3){
                        if($pol_sanger_seq_aligned[$i] eq "h"){
#                           print "\tiupac_agrees:\th\t$pol_sanger_seq_aligned[$i]";
                            $iupac_agreements++;
                        }else{
#                           print "\tiupac_disagrees:\th\t$pol_sanger_seq_aligned[$i]";
                            if($temp_alleles_array[1] eq $pol_sanger_seq_aligned[$i]){
                                $second_most_abundant_variant_agreements++;
                            }
                        }
                    }elsif($temp_alleles_string =~ /a/ && $temp_alleles_string =~ /c/ &&
$temp_alleles_string =~ /g/ && $temp_variant_count == 3){
                        if($pol_sanger_seq_aligned[$i] eq "v"){
#                           print "\tiupac_agrees:\tv\t$pol_sanger_seq_aligned[$i]";
                            $iupac_agreements++;
                        }else{
#                           print "\tiupac_disagrees:\tv\t$pol_sanger_seq_aligned[$i]";
                            if($temp_alleles_array[1] eq $pol_sanger_seq_aligned[$i]){
                                $second_most_abundant_variant_agreements++;
                            }
```

```
                                }
                        }elsif($temp_alleles_string =~ /a/ && $temp_alleles_string =~ /c/ &&
$temp_alleles_string =~ /g/ && $temp_alleles_string =~ /t/ && $temp_variant_count == 4){
                                if($pol_sanger_seq_aligned[$i] eq "n"){
#                                       print "\tiupac_agrees:\tn\t$pol_sanger_seq_aligned[$i]";
                                        $iupac_agreements++;
                                }else{
#                                       print "\tiupac_disagrees:\tn\t$pol_sanger_seq_aligned[$i]";
                                        if($temp_alleles_array[1] eq $pol_sanger_seq_aligned[$i]){
                                                $second_most_abundant_variant_agreements++;
                                        }
                                }
                        }else{
#                               print
"\tstill_does_not_agree:$temp_alleles_string\t$pol_sanger_seq_aligned[$i]";
                        }
                }
#               print "\n";
        }
}

#print "acgt_agreements:$acgt_agreements, iupac_agreements:$iupac_agreements, disagreements:$disagreements,
second_most_abundant_variant_agreements:$second_most_abundant_variant_agreements,
gap_disagreements:$gap_disagreements\n";
```

Appendix 3: Subtype and read abundance for each time point for each patient with multiple time points available. Percentages may not sum to 100% if not all reads mapped to the subpopulations present at >1%.

			Pol		Env	
Patient	Time point	Weeks post-diagnosis	COMET subtype	% abundance	COMET subtype	% abundance
Patient 1	T0	14	B	98.8	B	52.7
					B	17.9
					B	14.5
					B	11.0
	T1	27	B	96.0	B	53.9
			B	3.4	B	6.5
					B	3.3
					B	2.0
					B	3.0
					B	2.7
					B	5.7
					B	5.5
					B	8.0
	T2	39	B	97.7	B	20.8
					B	10.0
					B	3.3
					B	9.9
					B	2.7
					B	9.3
					B	19.9
					B	2.3
					B	6.2
					B	1.8
	T3	53	B	89.2	B	65.3
			B	7.8	B	2.1
					B	2.6
					B	6.8
					B	3.4
					B	3.4
					B	5.0
	T4	71	B	97.4	B	52.7
			B	2.5	B	3.3
					B	6.9
					B	3.0
					B	14.2
					B	3.0
					B	1.8
					B	8.5
Patient 2	T0	8	B	98.0	B	98.4
	T1	19	B	99.5	B	97.1
	T2	30	B	89.0	B	95.7
			B	5.5	B	2.7
	T3	33	B	99.2	B	96.8
					B	1.9

Patient	Time point	Weeks post-diagnosis	Pol		Env	
			COMET subtype	% abundance	COMET subtype	% abundance
Patient 4	T0	3	B	96.5	B	92.2
			B	1.7	B	6.0
	T1	8	B	57.6	B	76.7
			B	34.8	B	19.5
			B	2.5	B	2.4
	T2	11	B	54.5	B	67.7
			B	20.3	B	24.3
			B	19.5	B	6.0
			B	4.5		
	T3	19	B	56.9	B	40.6
			B	2.7	B	26.9
			B	31.3	B	19.3
			B	5.7	B	8.6
	T4	28	B	57.1	B	60.3
			B	20.9	B	10.0
			B	11.0	B	8.0
			D	1.9	B	9.3
					B	6.3
					B	3.4
	T5	33	B	49.5	B	50.1
			B	11.3	B	24.2
			B	34.0	B	2.4
					B	8.0
					B	2.1
					B	3.6
					B	1.8
					B	3.6
Patient 8	T0	21	B	80.5	B	96.2
			B	16.7	A1	2.5
	T1	26	B	90.8	B (check for 15_01B)	88.4
			B	2.4	A1	6.9
					A1 (check for 02_AG)	3.9
	T2	30	B	98.3	B (check for 15_01B)	68.7
					unassigned_1;D, A1	30.8

Patient	Time point	Weeks post-diagnosis	Pol		Env	
			COMET subtype	% abundance	COMET subtype	% abundance
Patient 9	T0	18	02_AG	95.4	A1 (check for 02_AG)	34.9
			02_AG	4.5	A1 (check for 02_AG)	2.1
					A1 (check for 02_AG)	7.9
					A1 (check for 02_AG)	7.0
					A1 (check for 02_AG)	1.7
					A1 (check for 02_AG)	1.8
					A1 (check for 02_AG)	4.7
					02_AG	6.8
					A1 (check for 02_AG)	3.3
					A1 (check for 02_AG)	3.3
					02_AG	5.3
					A1 (check for 02_AG)	5.3
					A1 (check for 02_AG)	2.0
	T1	27	02_AG	95.2	A1 (check for 02_AG)	45.1
					A1 (check for 02_AG)	2.0
					A1 (check for 02_AG)	3.4
					A1 (check for 02_AG)	8.7
					A1 (check for 02_AG)	5.5
					A1 (check for 02_AG)	5.5
					A1 (check for 02_AG)	6.0
					A1 (check for 02_AG)	1.6
					A1 (check for 02_AG)	1.8
					A1 (check for 02_AG)	2.6
					A1 (check for 02_AG)	2.8
					A1 (check for 02_AG)	3.2
					A1 (check for 02_AG)	3.0
	T2	40	02_AG	69.1	A1 (check for 02_AG)	53.1
			02_AG	23.2	02_AG	2.8
					A1 (check for 02_AG)	3.2
					A1 (check for 02_AG)	1.7
					A1 (check for 02_AG)	2.6
					A1 (check for 02_AG)	3.5
					A1 (check for 02_AG)	7.8
					A1 (check for 02_AG)	2.4
					02_AG	2.1
					A1 (check for 02_AG)	2.0
					A1 (check for 02_AG)	3.5
					A1 (check for 02_AG)	2.9
	T3	62	02_AG	92.0	A1 (check for 02_AG)	53.3
					A1 (check for 02_AG)	1.8
					A1 (check for 02_AG)	10.6
					A1 (check for 02_AG)	2.0
					A1 (check for 02_AG)	3.4
					A1 (check for 02_AG)	10.5
					A1 (check for 02_AG)	4.7

Patient	Time point	Weeks post-diagnosis	Pol		Env	
			COMET subtype	% abundance	COMET subtype	% abundance
Patient 13	T0	0	B	91.9	B	85.7
					B	3.2
					B	3.8
					B	7.0
	T1	1	B	84.9	B	73.8
			B	6.2	B (check for 15_01B)	12.8
			B	3.3	B	2.3
			B	2.6	B	8.1
					B	2.4
	T2	3	B	85.7	B	75.6
			B	3.7	B	4.3
			B	5.2	B	7.3
			B	4.1	B	4.1
					B	6.8
	T3	10	B	90.5	B	70.9
			B	3.0	B	5.4
			B	1.7	B	3.1
			B	1.7	B	7.2
					B	2.1
					B	3.1
	T4	14	B	82.3	B	38.8
			B	2.8	B	3.6
			B	4.7	B	9.6
			B	3.5	B	4.8
					B	3.2
					B	6.2
					B	8.3
					B	3.9
					B	1.8
					B	1.6
					B	2.5
					B	5.5
					B	1.9
	T5	16	B	83.5	B	31.8
			B	3.5	B	2.8
			B	3.5	B	3.3
			B	2.8	B	2.9
					B	1.8
					B	5.5
					B	12.8
					B	7.9
					B	4.5
					B	3.2
					B	5.8
					B	5.8

Patient	Time point	Weeks post-diagnosis	Pol		Env	
			COMET subtype	% abundance	COMET subtype	% abundance
Patient 23	T0	0	B	99.5	B	94.8
					B (check for 29_BF)	3.0
	T1	0	B	99.8	B	97.5
	T2	10	B	98.9	B	99.0
Patient 35	T0	0	B	99.1	B	99.2
	T1	1	B	97.3	B	95.0
	T2	3	B	99.8	B	99.2
Patient 36	T0	1	B (check for 29_BF)	99.4	B	99.9
	T1	7	B (check for 29_BF)	83.6	B	55.1
			B (check for 12_BF)	11.1	A1	44.2
	T2	25	B (check for 29_BF)	57.3	A1	75.3
			B (check for 29_BF)	3.2	F1	16.0
			B (check for 29_BF)	15.6	F1	5.1
			B (check for 29_BF)	3.4	A1	2.5
			D (check for 17_BF)	3.2		
	T3	37	B (check for 29_BF)	60.7	A1 (check for 02_AG)	94.1
			B (check for 29_BF)	19.5	B	5.0
			B	2.1		
			B	3.0		
			B	2.0		